TECHNICAL SPECIFICATION

ISO/TS 20281

First edition
2006-04-01

Water quality — Guidance on statistical interpretation of ecotoxicity data

Qualité de l'eau — Lignes directrices relatives à l'interprétation statistique de données écotoxicologiques

Reference number
ISO/TS 20281:2006(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In other circumstances, particularly when there is an urgent market requirement for such documents, a technical committee may decide to publish other types of normative document:

— an ISO Publicly Available Specification (ISO/PAS) represents an agreement between technical experts in an ISO working group and is accepted for publication if it is approved by more than 50 % of the members of the parent committee casting a vote;

— an ISO Technical Specification (ISO/TS) represents an agreement between the members of a technical committee and is accepted for publication if it is approved by 2/3 of the members of the committee casting a vote.

An ISO/PAS or ISO/TS is reviewed after three years in order to decide whether it will be confirmed for a further three years, revised to become an International Standard, or withdrawn. If the ISO/PAS or ISO/TS is confirmed, it is reviewed again after a further three years, at which time it must either be transformed into an International Standard or be withdrawn.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TS 20281 was prepared by Technical Committee ISO/TC 147, *Water quality*, Subcommittee SC 5, *Biological methods*.

Introduction

Ecotoxicity tests are biological experiments performed to examine if either a potentially toxic compound, or an environmental sample (e.g. effluent, sediment or soil sample) causes a biologically important response in test organisms. If so, the goal is to determine the concentration that produces a given level of effects or produces an effect that cannot be distinguished from background variation.

In a test, organisms are exposed to different concentrations or doses of a test substance or a test substrate (e.g. waste water, sludge, or a contaminated soil or sediment), sometimes diluted in a test medium. Typically, at least one group of test organisms (the control group) is not exposed to the test substance or substrate, but is otherwise treated in the same way as the exposed organisms.

The endpoint(s) observed or measured in the different batches may be the number of surviving organisms, size or growth of organisms, number of eggs or offspring produced or any relevant biochemical or physiological variable that can be reliably quantified. Observations are made after one or several predefined exposure times. The endpoint's relationship with the concentration of the tested chemical or substrate is examined. The way statistics are applied may have a considerable impact on the results and conclusions from ecotoxicity tests, and consequently on the associated policy decisions. Various documents (Williams 1971, Piegorsch and Bailer 1997, Tukey *et al*. 1985, Pack 1993, Chapman *et al*. 1995, Hoekstra 1993, Kooijman and Bedaux 1996, Laskowkj 1995, Chapman 1996, OECD 1998, ASTM 2000) exist on the use of available statistical methods, the limitations of these methods and how to cope with specific problematic data. Discussions of statistical principles and commonly used techniques are found in general references such as Armitage and Berry (1987) [basic information on hypothesis testing and regression, transformations], Finney (1978) [analysis of quantal data, especially probit models], Hochberg and Tamhane (1987) [thorough treatment of multiple comparison methods], Newman (1994) [information related to biology-based models, EC_x], and Sparks (2000) [a collection of articles covering field and laboratory experiments, multivariate techniques, risk assessment, and environmental monitoring].

When problematic data are encountered or critical decisions depend upon inferences from ecotoxicity tests, consultation with a qualified statistician is useful. (Statisticians should be consulted before beginning the experiment to ensure proper design, sample size, limitations, and to be sure that the study is actually able to answer the research question that the experimenter poses. Once bad data have been collected, there is little a statistician can do to rectify the problem.)

Clause 8 contains a table listing all the existing ISO and OECD ecotoxicity standards/guidelines that could be analysed using this guidance document. For each standard/guideline, reference is made to the adapted clauses of this Technical Specification.

Clause 4 details the different statistical approaches that can be used for the analysis of ecotoxicity data, depending on the aim. In particular, it gives the assumptions made when using hypothesis-testing methods, concentration-response modelling methods or biology-based methods and their limitations. It also gives some indication on experimental design issues. Some general principles and advice are also given for the process of data analysis.

Clause 5 deals with hypothesis testing, Clause 6 with dose-response modelling and Clause 7 with biology-based methods.

There was an ISO resolution (ISO TC 147/SC 5/WG 10 Antalya 3), as well as an OECD workshop recommendation (OECD 1998), that the NOEC should be phased out from International Standards.

However, the NOEC is still required in many regulatory standards from many countries, and in some cases, where a detailed determination of an EC_x is not relevant and the alteration of the study design is too costly to fulfil the requirements for regression models. Therefore guidance is provided on the statistical methods for the determination of the NOEC.

In the annexes, examples of analyses with the three main methods (hypothesis testing for NOEC estimation, dose-response modelling and biology-based modelling) of four different data sets are given. They concern:

— acute toxicity on *Daphnia magna*;

— inhibition of algae growth;

— reproduction of *Daphnia magna*; and

— fish growth.

Water quality — Guidance on statistical interpretation of ecotoxicity data

1 Scope

This Technical Specification offers guidance on statistical methods used for the analysis of data of standardized ecotoxicity tests. It focuses on statistical methods for obtaining statistical estimates of parameters in current and future use, e.g. EC_x (LC_x), NOEC, NEC.

The methods described are intended to cover laboratory ecotoxicity tests (aquatic, sediment and/or terrestrial tests), and may also be relevant for other toxicity tests.

The main objective of this Technical Specification is to provide practical guidance on how to analyse the observations from ecotoxicity tests.

Hypothesis testing, concentration-response modelling and biology-based modelling are discussed for the different data types (quantal, continuous and discrete data, corresponding to mortality, growth or reproduction).

In addition, some guidance on experimental design is given. Although the main focus is on giving assistance to the experimentalist, a secondary aim is to help those who are responsible for evaluating toxicity tests. Finally, the document may be helpful in developing new toxicity test guidelines by giving information on experimental design and statistical analysis issues.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3534-1: —[1)], *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

3 Terms and definitions

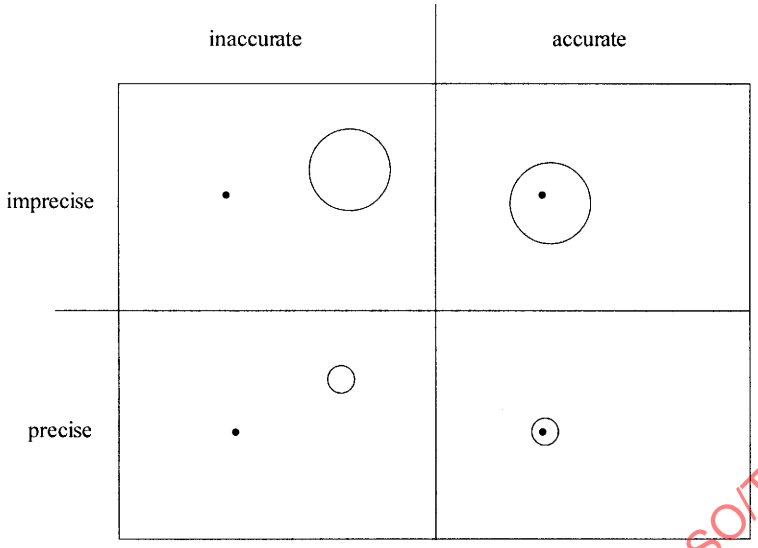
For the purposes of this document, the following terms and definitions apply.

3.1
accuracy
measure of how close the estimate is to the “true value” of the parameter (this true value is unknown)

3.2
precision
measure of the amount of variability in the estimate (quantified by the standard error or the confidence interval of the estimate)

1) To be published. (Revision of ISO 3534-1:1993)

NOTE Precision may be increased by using larger sample sizes or by reducing the experimental variation. However, as Figure 1 illustrates, although an estimate may be precise, this does not imply that it is also accurate.



NOTE The dot represents the true parameter value. The circle represents the confidence interval of the estimate. Small circles indicate high precision. Large circles indicate low precision.

Figure 1 — Conceptual illustration of accuracy and precision

3.3
concentration
amount of test material in the testing environment

NOTE 1 It is expressed, for example, as milligrams per litre in water, and as milligrams per kilogram in soil and in food.

NOTE 2 Concentration and dose both refer to the amount of test material to which the test organism is subjected. Statistical methods for both types of studies are identical; however, interpretations are different. Although "concentration" is used throughout this Technical Specification, all the statistical methods presented here also apply to studies in which a dose is used.

3.4
dose
amount of test material administered to a subject

NOTE 1 It is expressed, e.g. in milligrams/kilogram-body mass in an avian bolus study.

NOTE 2 Concentration and dose both refer to the amount of test material to which the test organism is subjected. Statistical methods for both types of studies are identical; however, interpretations are different. Although "concentration" is used throughout this Technical Specification, all the statistical methods presented here also apply to studies in which a dose is used.

3.5
confidence interval
x % confidence interval for a parameter is an interval of values that theoretically covers the true value of the estimated parameter with x % confidence

NOTE 1 Standard confidence intervals are based on the assumption that the underlying mathematical model is correct. It does not capture model uncertainty.

NOTE 2 A more precise definition is the following: interval estimator (T_0, T_1) for the parameter θ with the statistics T_0 and T_1 as interval limits and for which it holds that $P[T_0 < \theta < T_1] \geqslant 1 - \alpha$. [ISO 3534-1,—].

NOTE 3 Associated with this confidence interval is the confidence level 100(1 – α) % where α is generally a small number. The confidence level is typically 90 % or 95 %. The inequality $P[T_0 < \theta < T_1] \geqslant 1 - \alpha$ holds for any specific but unknown population value of θ.

NOTE 4 A confidence interval does not reflect the probability that the observed interval contains the true value of the parameter (it either does or does not contain it). The confidence reflects the proportion of cases that the confidence interval would contain the true parameter value in a long series of repeated random samples under identical conditions.

3.6 Data types

3.6.1
quantal/binary data
data that arise when a particular property is recorded to be present or absent in each individual

NOTE 1 An individual shows an effect or it does not show an effect. Therefore, these data can exhibit only two states.

NOTE 2 Typically, quantal data are presented as the number of individuals showing the property (e.g. mortality) out of a total number of individuals observed in each experimental unit. Although this can be expressed as a fraction, it should be noted that the total number of individuals cannot be omitted.

3.6.2
continuous data
data that can (theoretically) take any value in an open interval, for instance any positive number

EXAMPLES Measurements of length or body mass.

NOTE 1 Due to practical reasons, the measured resolution depends on the quality of the measurement device. For example, if test units are observed once per day, then “time to hatch” can only be recorded in whole days; however, the underlying distribution of “time to hatch” is continuous.

NOTE 2 Typically, continuous data have a dimension (e.g. grams, moles/litre).

3.6.3
discrete data
data that have a finite or countable number of values

NOTE There are three classes of discrete data: nominal, ordinal and interval.

— *Nominal data* express qualitative attributes that do not form a natural order (e.g. colours).

— *Ordinal data* reflect the relative magnitude from low to high (e.g. an individual shows no effect, minimal effect, moderate effect or high effect). These data cannot be interpreted with regard to relative scale (i.e. an ordinal variable with a value of “4” can be interpreted as being higher then the value of “2”, but not twice as high). Ordinal data can often be reduced to quantal data.

— *Interval* data (e.g. number of eggs or offspring per parent) allows the ranking of the items that are measured, and the differences between individuals and groups can be quantified. Often, interval data can be analysed as if the data were continuous. The analyses for interval discrete data are presented in this Technical Specification; analyses of nominal and ordinal data are not included but are addressed in a future revision.

3.7
effect
change in the response variable under consideration compared to a control

NOTE 1 For quantal endpoints, an effect is usually described in terms of a change in the percentage of individuals affected.

NOTE 2 For continuous endpoints, an effect is typically described in terms of a percent change in the mean values of the endpoint, but it can also be described in terms of absolute change.

3.8 Effect concentrations

3.8.1
quantal effective concentration
quantal EC_x
concentration of test material in water, soil, or sediment that causes x % change in response (e.g. immobility) during a specified time interval.

EXAMPLE An example of a concentration-response relationship and its associated estimates of EC_{10} and EC_{50}, are illustrated in Figure 2.

NOTE 1 It is expressed, e.g. in milligrams per litre or milligrams per kilogram.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.2
quantal effective dose
quantal ED_x
dose of test material that causes x % change in response (e.g. immobility) during a specified time interval

NOTE 1 It is expressed, e.g. in milligrams per kilogram body mass in an avian bolus study.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.3
quantal lethal concentration
quantal LC_x
concentration of test material in water, soil, or sediment that causes x % change in response (i.e. mortality) during a specified time interval.

EXAMPLE An example of a concentration-response relationship and its associated estimates of EC_{10} and EC_{50}, are illustrated in Figure 2.

NOTE 1 It is expressed, e.g. in milligrams per litre or milligrams per kilogram.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.4
quantal lethal dose
quantal LD_x
dose of test material that causes x % change in response (i.e. mortality) during a specified time interval

NOTE 1 It is expressed, for example, in milligrams per kilogram body mass in an avian bolus study.

NOTE 2 The x % change in response corresponds to an effect predicted on x % of the test organisms at a given concentration. This parameter is estimated by concentration-response modelling.

3.8.5
continuous effective concentration
continuous EC_x
concentration of test material in water, soil, or sediment that causes x % in the size of the endpoint during a specified time interval

NOTE 1 It is expressed, for example, in milligrams per litre or milligrams per kilogram.

NOTE 2 This parameter is also estimated by dose-response modelling.

3.8.6
continuous effective dose
continuous ED_x
dose of test material (e.g. mg/kg-body mass in an avian bolus study) that causes x % in the size of the endpoint during a specified time interval

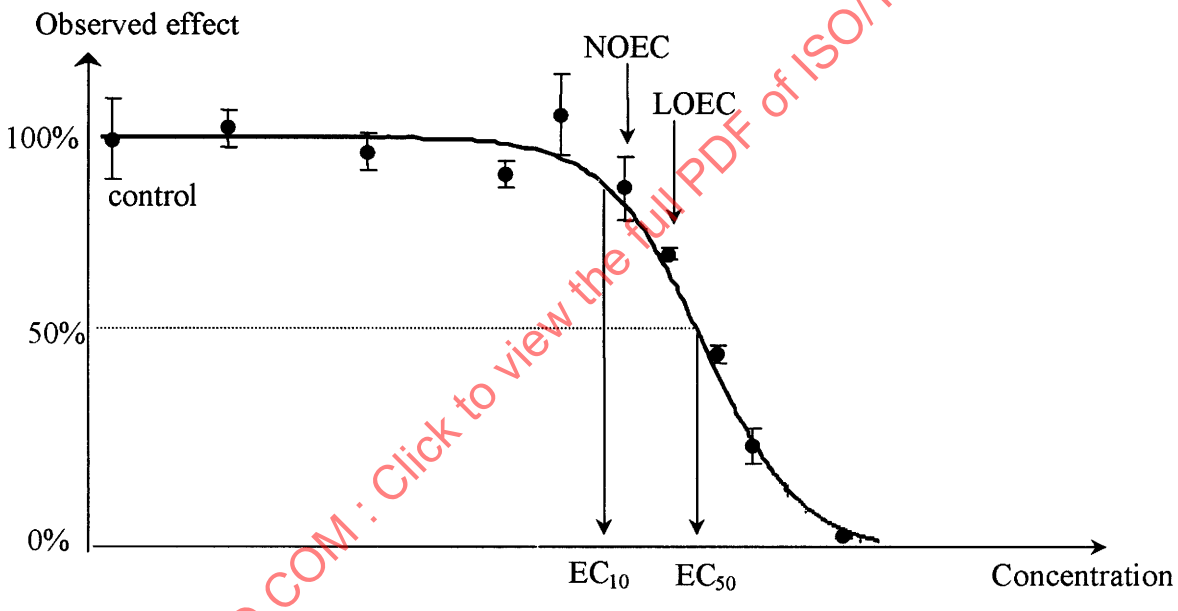
NOTE 1 It is expressed, for example, in milligrams per kilogram body mass in an avian bolus study.

NOTE 2 This parameter is also estimated by dose-response modelling.

3.9
endpoint
response variable
biological parameter observed

EXAMPLE Survival, number of eggs, size or growth, enzyme level.

NOTE An ecotoxicological study can have one or many endpoints.



NOTE The order of the parameters given in this figure has been taken at random.

Figure 2 — Illustration of a concentration-response relationship and of the estimates of the EC_x and NOEC/LOEC

3.10
effective time
ET_x
time at which an effect of x % is expected at a specified test concentration when the test organisms are exposed to a given concentration of material in water or sediment or soil

NOTE ET_x is estimated by modelling a time-response relationship.

3.11
lethal time
LT_x
time at which an effect of x % is expected at a specified test concentration when the test organisms are exposed to a given concentration of material in water, sediment or soil, and when the response of interest is mortality

3.12
experimental unit
replicate unit
smallest unit of experimental material to which a treatment can be allocated independently of all other units

EXAMPLES Aquariums, beakers, or plant pots.

NOTE By definition, experimental units shall be able to receive different treatments. Each experimental unit may contain multiple sampling units on which measurements are taken. Within each experimental unit, sampling units may not be independent. However, in some special case situations, individual organisms (housed in common units) can be treated as the experimental units: these special cases require some proof or strong argument of independence of organisms.

3.13
sampling unit
unit for which the measurement is taken

EXAMPLES Fish, daphnia or plants.

NOTE The sampling unit is not always identical with the statistical unit.

3.14
extrapolation
prediction of the value of variates outside the range of observations

NOTE 1 Extrapolation may not lead to a reliable estimate (see e.g. 6.4).

NOTE 2 When an EC_x estimated from a fitted concentration-response function is lower than the lowest nonzero concentration tested in the study, or higher than the highest concentration tested in the study, it is obtained by extrapolation.

3.15
interpolation
prediction of the value of variates within the range of observations

NOTE When the EC_x is between two consecutive nonzero test concentrations, it is said to be obtained by interpolation.

3.16
hormesis
effect where the tested substance is a stimulant in small concentrations, but it is inhibitory in large concentrations, resulting in a biphasic (or U-shaped) concentration-response relationship

NOTE 1 This observed stimulatory effect may be due to the tested substance, but it could also be due to an experimental artefact (e.g. solvent effect, non-random allocation of treatments to experimental units, experimental error).

NOTE 2 Models incorporating hormesis are not detailed in this Technical Specification. Two issues of *Critical Reviews in Toxicology* [2001, **31**(4) and **31**(5), pp. 351-694] and other journal articles discuss the issues concerning hormesis. Some discussion can be found in *Environment Canada* (2003).

3.17
lowest observed effect concentration
LOEC
lowest concentration out of the tested concentrations at which a statistically-significant difference from the control group is observed

EXAMPLE An example of LOEC is illustrated in Figure 2.

NOTE The LOEC is obtained by hypothesis testing.

3.18
no observed effect concentration
NOEC
tested concentration just below the LOEC

EXAMPLE An example of NOEC is illustrated in Figure 2.

NOTE The NOEC is obtained by hypothesis testing.

3.19
monotonic concentration-response
relationship in which the true, underlying concentration-response relationship exhibits an increase or a decrease over the range of concentrations in the study

NOTE If the concentration-response is monotone and non-increasing, the location parameters (mean or median) would exhibit the following relationship: $\gamma_0 \geqslant \gamma_1 \geqslant \gamma_2 \geqslant \gamma_3 \geqslant \ldots \geqslant \gamma_k$, where γ is the location parameter and 0, 1, 2, …, k are the concentration groups. If the monotone relationship is non-decreasing, the inequalities are reversed.

3.20
non-monotonic concentration-response
relationship in which the inequalities are not consistent across the concentrations

3.21
no effect concentration
NEC
parameter, which, when used in concentration-response models, is interpreted as being the highest concentration in which the compound does not affect the endpoint, even after very long exposure to the compound

NOTE The NEC equals the EC_0 at infinite time.

3.22
response
response corresponds to an observed value of any endpoint

3.23
parametric method
method which assumes that all the properties of the model are specified, except for the values of the parameters

EXAMPLE 1 In classical analysis of variance (ANOVA), the residuals are assumed to follow a normal distribution with a mean of zero and some unknown variance.

EXAMPLE 2 In Poisson regression, the response variable is assumed to follow a Poisson distribution (parameters to be estimated in the fitting process).

3.24
non-parametric method
method which contains weaker assumptions[2)] than the **parametric method** (3.23) about the shape of the distribution of the residuals, and the analysis is often based on ranks of the observations

NOTE The non-parametric analogue to a two-sample t-test is the Mann-Whitney test. Non-parametric regression is often conducted using a variety of smoothing techniques.

2) With fewer constraints.

3.25
parsimony principle
principle that data should be described with as few parameters as possible

NOTE A common decision criterion of including more parameters in the model is the observation that such leads to a significantly better description of these data.

3.26
systematic error
situation that a single concentration (dose) group differs from the others not only with respect to the intended treatment (i.e. the concentration or dose) but also with respect to some unintended experimental factors

EXAMPLE Containers housing the animals may differ by themselves, and in a design with few or only one container per dose group a deviating container may lead to a systematic error in that group.

NOTE The factor of time may underlie systematic errors in various ways, e.g. time of feeding, time of observation. The problem of systematic errors is that they may be wrongly interpreted as an effect of the intended treatment.

3.27
statistical significance
observed deviation from what was expected under the null hypothesis, it is unlikely to be attributable to chance variation

EXAMPLE In hypothesis testing, a result is statistically significant at the chosen level α if the test statistic falls in the rejection region.

NOTE In this Technical Specification, the α-level is 0,05 unless otherwise stated.

4 General statistical principles

4.1 Different statistical approaches

4.1.1 General

For each of the three analysis methods introduced below, it is necessary to obtain data from a designed experiment with replications of controls and concentration groups. All three classes of analysis methods (hypothesis testing, concentration-response modelling and biology-based methods) are suitable for data from toxicity tests as currently standardized by several OECD and ISO guidelines. However, designs for each of these studies can be optimized with respect to cost-effectiveness and the selected analysis approach. The number and spacing of the concentrations depend on the study being conducted and the type of data analysis to be utilized.

For each of the three approaches introduced below, the following is provided:

— a brief description of the use of each method in ecotoxicity tests;

— a brief outline of specific analysis methods presented in the later clauses of this Technical Specification;

— a listing of some major assumptions and limitations for each approach.

4.1.2 Hypothesis-testing methods

Hypothesis testing is a statistical inference technique used to compare the responses among two or more test groups. Hypothesis testing has many uses in ecotoxicology, ranging from detecting whether there is a significant difference in the measured response between the control and a given concentration, to establishing an LOEC and an NOEC. Discussion in this Technical Specification focuses on use in determining LOECs and NOECs, the most frequent use of hypothesis testing in OECD guidelines.

Methods discussed in Clause 5 include analyses for quantal data and continuous data. For both types of data, parametric approaches (when an underlying distribution e.g.: normal, log-normal is characterised) and non-parametric approaches (when weaker assumptions are made regarding the distribution) are presented. In Clause 5, assessment is limited to conducting data analysis separately at each time point, though this is not a limitation of the method. Three terms often used when discussing hypothesis tests are: Type I errors, Type II errors, and power (Table 1).

— *Type I errors* (false positives) occur when the null hypothesis is the truth but the hypothesis test results in a rejection of the null hypothesis in favour of the alternative hypothesis. The probability of making a Type I error is often referred to as α and is usually specified by the data analyst – often at 0,05, or 5 %.

— *Type II errors* (false negatives) occur when the alternative hypothesis is true but the test fails to reject the null hypothesis (i.e. there is insufficient evidence to support the alternative hypothesis). The probability of making a Type II error is often referred to as β (1 – power).

— *Power* is the probability of rejecting the null hypothesis (H_0) in favour of the alternative hypothesis (H_A), given that the alternative hypothesis is true. Power of a test varies with sample size, variance of the measured response, the size of an effect that it is of interest to detect, and the choice of statistical test. Power to detect differences can be increased by increasing the sample size and/or reducing variation in the measured responses. Thus, if a test has low power to detect an effect of a given size, this is equivalent to saying that the test has a low probability of detecting an effect of that size.

Table 1 — Probabilities of finding a significant or non-significant test outcome, given that the null hypothesis is true or not

Result of hypothesis test	State of the world	
	H_0 true	H_A true
not significant	$1 - \alpha$	Type II error β
significant	Type I error α	$1 - \beta$ = power

Several assumptions made when conducting hypothesis tests to determine the NOEC are given below.

— The concentration-response relationship may or may not be assumed depending on the specific statistical tests used.

— This approach makes only weak assumptions about the mechanisms of the toxicant or the biology of the organism.

Several limitations of using hypothesis testing to determine the NOEC are given below.

— Since the NOEC (or NOEL) does not estimate a model parameter, a confidence interval cannot be assessed.

— The value of the NOEC is limited to being one of the tested concentrations (i.e. if different values were chosen for the tested concentrations, the value of the NOEC would be different).

— If the power is low (due to high variability in the measured response and/or to small sample size), the biologically important differences between the control and treatment groups may not be identified as significantly different. If the power is high, it may occur that biologically unimportant differences are found to be statistically significantly different.

4.1.3 Concentration-response modelling methods

Regression methods are used to determine the relationship between a set of independent variables and a dependent variable. For designed experiments in ecotoxicology, the main independent variable is the concentration of the test substance and the dependent variable is the measured response (e.g. percent survival, fish length, growth rate). Regression methods fit a concentration-response curve to the data and use this curve to estimate an Effective Concentration (EC_x) at a given time point. The mathematical model used may be any convenient function that is able to describe the data; however, some models are more frequently used and accepted within the ecotoxicity testing literature. Several methods are available for model fitting and parameter estimation.

Methods discussed in Clause 6 include analyses for quantal data and continuous data. Parametric approaches (when a specific underlying distribution is assumed) are presented.

Although non-parametric methods have been developed for fitting concentration-response curves and estimating an EC_x, they are not presented in this Technical Specification.

— Sources on non-parametric regression include:

  Green and Silverman (1994), Easton and Peto, Fan and Gibjels (1996), Hardle, W. (1991), Azzalini and Bowman (1997), Silverman B. (1985), Akritas and Van Keilegom (2001), Carroll *et al.* (1999) and Smith-Warner *et al.* (1998).

— Software for non-parametric regression can be found.

The effect of exposure time is considered in Clause 6.

Although power is typically only discussed when hypothesis tests are conducted, both sample size and variation in the response variable within groups affect the inferences of concentration-response models as well. Small sample sizes and high variability in the response within groups increase the width of the confidence interval of the parameters of interest (e.g. EC_x), and the fitted model may not reflect the true concentration-response relationship. To increase the level of confidence in the parameter estimate, the number of replicates can be increased and measures to minimize unexplained variability could be taken. The width of the confidence interval also depends on the experimental design (i.e. the location and number of concentrations chosen). Finally, in addition to precision, accuracy of the estimated parameter is just as important. To enhance accuracy, concentrations should be chosen such that various different response levels are observed.

These specific properties of the experimental design, the number and spacing of doses and the number of replicates, are related to the value of x of interest in the particular experiment. Different designs may be employed to estimate an EC_{50}, as opposed to an EC_{05}. Further guidance for the design of experiments of this type is discussed in Clause 6.

Several assumptions of concentration-response modelling are:

— the models discussed in this Technical Specification assume the responses have a monotonic concentration-response relationship;

— the fitted curve is close to the true concentration-response relationship;

— this is an empirical model and does not make strong assumptions about the mechanisms of the toxicant or the biology of the organism.

Several limitations of concentration-response modelling are:

— estimation of EC_x values outside the concentration range introduces a great deal of uncertainty (i.e. extrapolation outside the range of the data);

— once the experiment has been performed, the resulting concentration-response data may not be suitable for the estimation of parameters of a concentration-response model. In particular, when the gaps between consecutive response levels are so large that many different concentration-response models would fit equally well to the observed data, interpolation would not be warranted.

4.1.4 Biology-based methods

The biology-based methods presented in this Technical Specification provide models for exploring the effect of the test chemical over time as well as incorporating a toxico-kinetic model for the behaviour of the chemical. By modelling concentration and exposure time simultaneously, these methods fit response surfaces to response data to estimate an EC_x as a function of exposure time, rather than fitting separate response curves at each time point.

Methods discussed in Clause 7 include analyses for quantal data and continuous data for several aquatic toxicity tests (acute and chronic tests on survival/immobility for daphnia and fish, fish growth test, daphnia reproduction test, and alga growth inhibition test). The models presented in this Technical Specification utilize dynamic energy budget theory (see Clause 7 for details and associated references). This theory provides a quantitative description for the processes of feeding, digestion, storage, maintenance, growth, development, reproduction and ageing and their interrelationships. As with concentration-response modelling, the level of confidence in the parameter estimates (as evidenced by the width of the confidence interval) is a function of sample size and inherent variation in the response.

Because of additional assumptions regarding the toxico-kinetic behaviour of the chemical and the biological behaviour of the organism in the system, it is sometimes possible to carry out additional extrapolation from the toxicity test. The assumptions are endpoint-specific; therefore, for each type of test, these assumptions need to be defined. The definition of these assumptions usually involves eco-physiological background-research prior to the specification of the test. However, if these additional assumptions can be made, examples of additional outcomes of this method can predict, e.g. chronic responses from acute responses, responses to time-varying concentrations using responses to constant concentrations, and responses by a species using responses to a con-specific or physiologically-related species of a different body size for a given test compound.

Several general assumptions made when using biology-based methods are the following.

— The models discussed in this Technical Specification assume the response has a monotonic concentration-response relationship.

— This analysis method incorporates mechanistic models for toxico-kinetics and physiology.

Several limitations of biology-based methods are the following.

— Estimation of parameter values (e.g. EC_x and NEC) outside the concentration range introduces a great deal of uncertainty (i.e. extrapolation outside the range of the data).

— When the gaps between consecutive response levels are so large that different biology-based models would fit equally well to the observed data, NEC estimations would not be warranted, if they differ substantially between the models.

— To date, models have been developed for some of the common aquatic toxicity tests (acute and chronic tests on survival/immobility for daphnia and fish, fish growth test, daphnia reproduction test, and alga growth inhibition test). Nevertheless, these models can be applied to any test species.

4.2 Experimental design issues

4.2.1 General

The usual factors (independent variables) studied in ecotoxicity tests are the concentration of the tested substance and the duration of exposure. For the estimation of the effect at a given condition, it is necessary to replicate these conditions, to control experimental variation (see 4.2.3).

The experimental design, amongst others, specifies the tested concentrations of the substance, the number of replicates and the number of containers per tested concentration, as well as the times of observation.

4.2.2 NOEC or EC_x: Implications for design

The estimation of an EC_x puts different demands on the study design than does the assessment of an NOEC.

— When the aim is to assess an NOEC, an important demand is that the study warrants sufficient statistical power. To that end, the concentration (dose) groups need a sufficient number of replicates (possibly at the expense of the number of dose groups). Many test guidelines are based on this principle.

— When the aim is, however, to provide an estimate of an EC_x, the primary demand on the study design is to have a sufficient number of concentration (dose) groups. This may be at the expense of the number of replicates per group (e.g. keeping the total size of the experiment the same), since the precision of the estimated EC_x depends more on the number and spacing of concentrations than on the sample size per concentration or dose group.

— The demands for study designs aimed at estimating an NOEC or an EC_x are further discussed in 5.1.7 and 6.5, respectively.

Therefore, the choice between assessing an NOEC and estimating an EC_x should actually be made before designing the study. If one wishes (or is required) to assess both, a compromise between the two opposing demands shall be made, i.e. a design with both a sufficient number of dose groups and a sufficient number of replicates in each group.

— The number of replications per group needed for assessing an NOEC depends on the desired power of the statistical test involved (see 4.1.1).

— For assessing an EC_x, three concentration groups, next to the control group, is an absolute (theoretical) minimum. However, when just one dose group appeared to have been unluckily chosen, the assessment of an EC_x would probably fail, and more concentration groups are therefore required in practice.

Design recommendations for experiments using a biology-based model include those for EC_x. Additional recommendations are discussed in 7.8.3.

4.2.3 Randomization

Variability is inherent in any biological data set. This variability is directly visible in continuous and discrete data. Although the following discussion holds for any type of data, it is easiest to use continuous data as an example. In analysing concentration-response data by statistical methods, the observed scatter is sometimes called noise or variation, but when designing experiments and interpreting results it is good to understand the reasons for the noise. The following factors may play a role:

a) the variation between the individual animals, due to genetic differences;

b) the differences in the conditions under which the animals grew up prior to the experiment, resulting in epigenetic differences between animals;

c) the heterogeneity of the experimental conditions among the animals during the experiment;

d) variation within subjects (i.e. fluctuations in time, such as female hormones, which may be substantial for some endpoints); and

e) measurement errors.

Randomization processes are used in designed experiments to eliminate bias in estimates of treatment effects, and to ensure independence of error terms in statistical models. Ideally, randomization should be used at every stage of the experimental process, from selection of experimental material and application of treatments, to measurement of responses.

To minimize the effects of the first two factors, animals need to be randomly distributed into concentration groups. To minimize the effects of the third factor (both intended and unintended, such as location in the room), application of treatments should be randomized as much as possible. To minimize the effects of the fourth factor, the measurement of responses should be randomized in time (e.g. although all responses are recorded at 24 h, the order in which the experimental units are measured should be randomized). With good scientific methods, measurement errors can be minimized.

In some circumstances, it may be difficult, or expensive, to randomize at every stage in an experiment. If any experimental processes are carried out in a non-random way, then statistical analysis of the experimental data should include a phase in which the potential effect of not randomizing on the experimental results is examined.

4.2.4 Replication

As discussed above, noise cannot be avoided, and therefore it is necessary to assign a certain number of replicates (experimental units) to each treatment group and control group. The number of replicates influences the power in hypothesis testing and the confidence limits of parameter estimates. A standard assumption of all methods is that replicates are independent. Treating observations as independent replicates, whereas in fact they are not, represents an error called pseudo-replication (Hurlbert, 1984). This issue becomes important when animals are housed together, as in a tank or beaker.

There are two types of housing effects:

— containers may differ from each other in some (usually unknown) sense;

— the organisms within a container affect each other's responses.

Both effects result in non-independence (or pseudoreplication) of the individual organism's responses.

The first effect may result in different mean responses between containers (at a given concentration). This type of non-independence can be addressed by taking the variation between containers into account in the statistical model. For instance, with continuous data this may be done using a nested ANOVA, where the individual observations are nested within the container.

The second effect might distort the distribution of the observations related to the individual organisms. For instance, with quantal data, the assumption of binomial distribution may not hold. In an example with continuous data, when there is competition among individuals in the same container, the responses of the individual organisms may appear bimodal. See Clause 5 and Clause 6 for more detailed discussions.

4.2.5 Multiple controls included in the experimental design

It is common in aquatic and certain other types of experiments that the chemical under investigation cannot be administered successfully without the addition of a solvent or vehicle. In such experiments, it is customary to include two control groups. One of these control groups receives only what is in the natural laboratory environment (e.g. dilution water), while the other group receives the dilution water with added solvent but no test chemical. In ecotoxicity experiments, these are often termed negative or dilution water (non-solvent) and solvent controls.

OECD recommends limiting the use of solvents (OECD 2000); however, appropriate use of solvents should be evaluated on a case-by-case basis. Details regarding the use of solvents (e.g. recommended chemicals, maximum concentrations) are discussed in the relevant guideline documents for a specific ecotoxicity test. In addition, regulatory guidelines shall be followed by both controls with regard to the range of acceptable values (e.g. minimum acceptable percent survival or mean oyster shell deposition rate). Multiple control groups can be utilized regardless of whether the experiment was intended for hypothesis testing (i.e. determination of an NOEC), regression analysis (i.e. determination of an EC_x), or biology-based methods. The focus of this subclause is to present a data analysis methodology for experiments in which a solvent is used.

Data from the two control groups are analysed to determine if the solvent had a statistically significant effect on the measured response. If there is a statistically significant difference between the negative and solvent

control groups, any conclusions and inferences based on this study could be impacted due to the presence of a solvent effect. If there are no significant differences in the means (or proportions for quantal data or medians for non-parametric data) between the negative and the solvent controls, then it is concluded that there is insufficient evidence to detect a difference between the controls.

The solvent control group is the appropriate control group for comparisons with treated groups. Each group shall have the same solvent concentration as the control. For a toxicity test in which a solvent is used in conjunction with the test chemical, the assumptions are that the solvent had no effect on the responses of interest and there was no interaction between the test chemical and the solvent. With the addition of a negative control (as is required in all experiments using a solvent), the assumption regarding a solvent effect can be tested. However, unless the chemical is also tested in absence of a solvent, the assumption of no interaction between the solvent and the test chemical cannot be evaluated.

Some practitioners consider combining the data into one “pooled control” for comparison to the treated groups when no statistically significant differences between the solvent and negative control were identified. However, this does not take into account the fact that the differences between the two controls might not have been detected with a statistical test because the sample sizes are too small (i.e. low power) or that it combines two sources of variability.

The methods used for statistical comparison of negative and solvent controls vary depending on the type of data and the assumptions regarding distribution of the data. Methods and mathematical details for carrying out these tests are found in Clause 5 and its associated annexes.

4.3 Process of data analysis

4.3.1 General

A typical data analysis more or less follows a general pattern, usually constituting the following steps. First, the data are plotted and visually inspected. Then, a suitable type of analysis is chosen, based on particular assumptions that appear reasonable for the data at hand. After the analysis, the underlying assumptions are checked. If necessary, an adjusted analysis is performed. And finally, the results are reported by making plots and/or tables.

4.3.2 Data inspection and outliers

A useful first step in analysing dose-response data is to visually inspect the data. For continuous data, the individual responses (together with the group means) may be plotted as a function of dose. For quantal data, one may plot the observed frequencies of response as a function of dose. These plots are useful to assess whether the data show a dose-response relationship. Further, these plots may indicate any peculiarities in the data. In particular, the observed data may show outliers, i.e. data points far away from intuitive expectation, or from the general pattern shown by the data. In continuous data, one may detect both outliers that relate to the individual organism (or, more generally, the biological system serving as the experimental unit), and outliers that relate to a whole treatment group. In quantal data, outliers always relate to a treatment group, since a deviating individual cannot be detected based on a “yes” or “no” response.

Outliers that relate to a whole treatment group may arise due to the fact that a treatment group differed systematically from the other groups by some (usually unknown) experimental factor(s). For instance, the organisms in the various dose groups were held in different aquaria, and one of them contained an infection. Or, the organisms in the different dose groups were treated in a specific order (with respect to feeding, time of observation, etc.). Detection of this type of outliers typically cannot be enforced by any formal statistical method, and one has to rely on visual inspection, judgement and experience.

Obviously, treatment group outliers are highly undesirable, since they directly interfere with the effect that one wishes to measure, thereby increasing the probability of both false positive and false negative results. For example, an NOEC may be assessed at a level where substantial effects do occur, or an LOEC may be assessed at a level without real effects (i.e. from the chemical). The only way to prevent outliers at the group level is a design that is perfectly randomized with respect to all experimental actions that may potentially influence the (observed response of) the biological system. In practical biological studies, however, perfect randomization is hard to realise, and it is not feasible to reduce the probability of getting group outliers to nil.

Therefore, it is paramount to make the study design relatively insensitive to potential outliers, i.e. by randomized replicated dose groups, and/or by increasing the number of different doses (followed by dose-response modelling, see Clause 6).

Outliers at the individual level can only be detected in continuous data. When a particular distribution is assumed for the scatter in the data, the judgement of outliers may be based on a specific, small probability that any single data point could occur. This implies that the judgement of outliers can depend quite strongly on the assumed distribution. For example, values that appear to be extremely high when assuming a normal distribution may be judged as non-extreme when assuming a log-normal distribution. Vice versa, low values may be judged as extremes when assuming a log-normal distribution, but not so when a normal distribution is assumed.

The statistical analysis of the data is sensitive to individual outliers, although less dramatically than to group outliers. On the one hand, individual outliers may result in biased estimates of the effect (either too small or too large). On the other hand, the estimate of the residual variance (the "noise") is increased, implying that statistical tests tend to be less powerful, and estimated parameters (e.g. EC_x) less precise. Therefore, if reasons can be found explaining the outliers, it is favourable to delete them from the analysis.

Although non-detectable, individual outliers can also occur in quantal data and affect the analysis. For example, when just one of the individuals in the controls shows a response, but is in fact an outlier, this outlier may have quite an impact on the statistical analysis. Being non-detectable, individual outliers are a larger problem in quantal data than in continuous data.

In conclusion, outliers can have dramatic effects on the statistical analysis and the conclusions drawn. Therefore, it is very important to reduce their impact by using designs that are relatively insensitive to them, i.e. by utilizing replicated dose groups and/or multiple dose groups. More information can be found in Atkinson (1985), Belsey *et al.* (1980) and Cook and Weisberg (1982).

4.3.3 Data inspection and assumptions

4.3.3.1 Scatter

Visual inspection may be used to explore the general pattern of the scatter around (continuous) data. Thus, one may find out if the scatter around the mean response appears to be symmetrical or skewed, and if the scatter is more or less homogenous. Heterogeneity of variance (scatter) may have a biological basis, i.e. the individual organisms (units) respond differently to the chemical. However, an apparent increase or decrease in the scatter may also be related to the statistical distribution of the data, e.g. the scatter increases with the mean response. This distinction is important, both for the analysis and for the interpretation of results. Further clarification is given below.

4.3.3.2 Heterogeneous variances and distribution

When the plotted data show scatter that is correlated with the mean response, such a pattern may be related to the underlying distribution of the data. Some examples illustrate this.

In log-normally distributed data, it may be theoretically expected that the standard deviation increases proportionally with the mean (or, equivalently, the Coefficient of Variation, CV, is homogenous). Also, for the gamma distribution, the CV is expected to be homogenous. When a particular dataset (such as masses, concentrations) shows scatter that increases with the mean, one may plot such data on the log-scale, which usually makes the scatter independent from the means. In addition, when the scatter is relatively large (say, with a CV larger than 20 %), the scatter may be skewed on the original scale, but not on the log-scale. The latter would confirm that the pattern in the scatter is a result of the underlying distribution.

As another example, counts may follow a Poisson distribution. Here, the variances are expected to be equal to the means (or proportional to them). Such a pattern should vanish when the data are plotted on square root scale.

Finally, in quantal data with replicated dose groups, it can be also be expected *a priori* that the scatter between the replicates depends on the mean response (this follows directly from the properties of

frequencies). Here, one may plot the frequencies after the transformation arcsin($\sqrt{p}$), where p is the observed frequency (fraction). This transformation is able to remove the (theoretical) relationship between the variance and the mean frequency (assuming a binomial distribution).

4.3.3.3 Heterogeneous variances and true variation in response

Heterogeneity in the scatter might also be caused by the treatment (the applied chemical) itself, i.e. some individuals respond stronger to the chemical than others. This could happen when genetically heterogeneous organisms are used, e.g. subject to genetic polymorphism. In many toxicity tests, however, the organisms used are genetically homogenous, and real (biological) heterogeneity in response to the chemical is, in those cases, not very likely.

4.3.3.4 Consequences for the analysis

Heterogeneity of variances may be a matter of scaling that can be removed by the right transformation. Usually such a transformation also tends to make the data more normally distributed. Thus, one may apply standard methods based on normality (e.g. t-test, ANOVA, linear regression) to the transformed data. Another approach is to omit the transformation, and use methods that are directly based on the assumed distribution (i.e. generalized linear models). When a particular transformation is found that results in homogenous variances, only one variance parameter needs to be estimated. Thus, all the data contribute in the variance estimate, which is in statistical terms reflected by a larger number of degrees of freedom [3)].

However, when the heterogeneity of variances appears to be due to real biological heterogeneity in responses among individual organisms, one should carefully consider if further analysis is meaningful. For example, when the organisms (or experimental units) consist of two distinct subpopulations, one responding, the other not, any estimated change in mean response has no useful meaning. When such two subpopulations can be distinguished from observable features (e.g. sex), the appropriate way to proceed is to analyse both subpopulations separately, or by using the observable feature as a covariate (see, e.g. 6.3.3, and Figure 14).

4.3.4 Transformation of data

Many standard parametric methods (e.g. ANOVA, t-tests, linear regression analysis) assume normally distributed data and homogenous variances. In practice, the data often deviate from these assumptions, and if so, the inferences resulting from these standard methods may be disturbed. A variance-stabilizing transformation is often applied to the data, and then the statistical analysis is performed on the transformed data. Examination of residual plots (plot of the residuals vs. the predicted values) and tests of heterogeneity of variance (e.g. Levene, Bartlett, Hartley's F-max, or Cochran's Q) can help to identify instances when the variances among the concentration groups are unequal. References on this topic include Box and Cox (1964), Box and Hill (1974), Box and Tidwell (1962), Draper and Cox (1969).

For a variance-stabilizing transformation to exist, there shall be a relationship between the population means and variances. In many cases, the theoretical distribution of the response variable can guide the choice of a transformation. Examples are given below.

— If the underlying distribution is assumed to be Poisson, the square root transformation,
$y_i' = (y_i^{1/2})$ or $y_i' = [(y + 1)^{1/2}]$, is used.

— If the underlying data are log-normal, the log-transformation, $y_i' = \log(y_i)$, is often used.

— For proportions with binomial distributions, the arc-sin square-root [$y_i' = \arcsin(y_i^{1/2})$] and
Freeman-Tukey [$y_i' = (y_i + 1)^{1/2} + (y_i^{1/2})$] transformations are often used.

3) In general, it is favourable to include as few parameters (that need to be estimated from the data) as possible in the analysis, and yet describe the data accurately. Too few parameters probably result in biased estimates; too many parameters tend to result in too wide confidence intervals. This is also referred to as the parsimony principle.

— If the underlying theoretical distribution is unknown, a data-based procedure (Box-Cox transformation) can be used (Box and Cox,1964).

The use of transformations often simplifies the data analysis, in that the more familiar and traditional data analysis methods can be used, but care shall be taken in interpreting the results of this data analysis. Several aspects are discussed below.

If a transformation is used, it is also necessary to back-transform the means and confidence intervals to the original scale, when reporting results. It is not correct to back-transform the standard errors. It is important to understand that the back-transformed means differ from the arithmetic means of the original data. These back-transformed means should be interpreted as estimates of the median of the underlying data distribution, if the transformed data are normally (or at least symmetrically) distributed. In the special case of a log-transformation, the back-transformed mean is the geometric mean of the original data, and this value estimates the median of the underlying log-normal distribution.

When a transformation is not used in the data analysis, the difference in the means is a logical measure for the size of an effect. This difference is interpreted as an absolute change in the original units (e.g. a decrease of 1,2 g). The back-transformed difference in means (of the transformed data) however has another, usually more difficult, interpretation. In the special case of a log-transformation, the difference between the back-transformed means does allow a simple interpretation: it estimates the ratio (or percent change) of the median responses.

In addition, transformations may not maintain additivity of effects (interactions among factors, e.g. test substance, sex and age in the experiment). Other possible consequences of using transformations are that they change the interpretation of outliers and that they affect the value of R (Pearson correlation coefficient) and R^2 (coefficient of determination). Not all data problems can be fixed by transformation of the response. For example, if a large percentage of the responses have the same measured value (ties), no transformation addresses that issue.

4.3.5 Parametric and non-parametric methods

4.3.5.1 General

A visual inspection of the data may have indicated that the scatter is more or less symmetric and homogeneous, possibly after a particular transformation. In that case, one may analyse the data by the standard parametric methods based on normality. Or, one may choose to analyse the data based on a particular distribution other than the normal. Here, some basic aspects of parametric and non-parametric methods are discussed.

4.3.5.2 Parametric methods

When the data are assumed to follow a particular statistical distribution, they can be summarized by the parameters of that distribution. For example, data that are normally distributed can be summarized by just two parameters, the mean and the variance. Therefore, methods that are based on an assumed distribution are called “parametric methods”. Obviously, these methods intend to estimate the parameters of the (assumed) distribution, such as the mean and the variance, or any derived parameters, such as the EC_x.

If one is interested in the value of some entity (such as the EC_x), rather than a “categorical” answer (significant or nonsignificant), parametric methods are the natural approach of analysis. In addition, in hypothesis testing, parametric methods such as ANOVA are also widely used.

Whatever distribution is assumed, parametric methods are based on the general principle of fitting the data to the model. In hypothesis testing, this may be the ANOVA model; in dose-response modelling, this may be a particular dose-response model.

— In applying parametric hypothesis tests, one shall examine the data for outliers, deviations from normality and homogeneity, assessment of monotonicity of the dose response (for some approaches), and do a general assessment of whether the proposed model adequately describes the data. These points are discussed in depth in 5.1, 5.1.4, 5.1.6, 5.2.2.3, 5.2.2.5, 5.3.1, 5.3.1.3, 5.3.1.4, 5.3.1.6, and 5.3.1.7.

— In dose-response modelling, the process of model fitting is eminent and indeed the focus of the analysis. Therefore, in any dose-response analysis (as discussed in Clauses 5, 6 and 7) the user should understand the general principles of model fitting. These are discussed in 4.3.6, 5.1.4, 5.2.2, 5.3.1, and 6.7.

4.3.5.3 Generalized linear models (GLMs)

Generalized linear models are an alternative approach to use parametric methods when the normality assumption is violated. In this approach, the analysis of the (untransformed) data is based on another (than normal) distribution, for example, a Poisson distribution (for counts), or a binomial distribution for frequencies. GLMs are not discussed in this Technical Specification, and the reader is referred to the literature (Mc Cullagh and Nelder, 1983; Kerr and Meador, 1996).

4.3.5.4 Non-parametric methods

Non-parametric methods have been developed for those cases where one is not willing to assume any distribution at all. These methods can be used to test the null hypothesis that the observations in two (or more) treatment groups do not differ (i.e. they stem from the same, but unknown distribution). These methods are based on the rank order of the observations. Therefore, significantly different treatment groups are supposed to differ in the medians (since the median can be defined in terms of rank order). To prevent misunderstanding, the medians should always be reported when non-parametric methods were used. (Differences between means may not be consistent with differences between medians, i.e. means are more sensitive to outliers than medians are).

4.3.5.5 How to choose?

Parametric analyses have various advantages over non-parametric methods.

— They are typically simpler to conduct (wide availability of software).

— The methods have been developed for a wider array of designs (e.g. designs with replicated dose groups).

— The confidence intervals are more easily computed.

— The methods are more universally used.

— Interpretation of results is often easier.

Non-parametric methods have the following advantages.

— They are based on very weak assumptions.

— Further, since non-parametric analyses are based on the rank order of the data, they are less sensitive to outliers than parametric analyses.

When the data appear to comply with the assumptions (after a visual inspection) of a particular parametric analysis, parametric is the obvious method to choose. The assumptions can be further checked as part of the analysis (e.g. by examining the residuals, see below). It may be noted that parametric analysis based on normal assumptions is reasonably robust to mild violations against the assumptions. When a data transformation results in a (better) compliance with the normality assumptions, one should be reminded that transformations other than the log-transformation may impair the interpretation of the results. This is because the log-transformation is naturally linked to the intuitive notion that biological effects are proportional (or multiplicative) rather than additive (compare definition of EC_x). Thus, when omitting or applying a log-transformation does not make a large difference for complying with the assumptions, one might choose to apply it for reasons of interpretation.

In situations where specific distributions are natural candidates for the data type at hand, one may consider the use of GLMs.

When no regular distributions can be assumed, as e.g. in the case of tied observations[4)], one may resort to non-parametric analysis.

4.3.6 Pre-treatment of data

In general, pre-treatment of data (other than data transformation) is not a favourable strategy for data analysis. A few practical examples are discussed.

Some methods (e.g. the probit and logit models for quantal dose-response analysis) use a log-transformation for concentration. It is not appropriate to add a small positive constant to the zero-concentration (or to all concentrations) to avoid taking the log of zero (see Clause 6 for more details). The shape of the concentration-response curve is very sensitive to the constant and a biological basis for choosing one constant over another is very unlikely to ever exist.

A current habit in analysing continuous data is to divide the observed responses by the (mean) observed response in the controls. These corrected observations then reflect the percent change compared to the controls, which is usually the entity of interest. However, such a pre-treatment of the data is improper. Among other problems, it assumes that the (mean) response in the controls is known without error, which is not the case. Therefore, this should be avoided, and instead the background response should be estimated from the data by fitting the model to the untreated data. Thus, the estimation error in the controls is treated in the same way as the estimation errors in the other concentration groups. (see e.g. 6.2.3 and 6.3.3).

4.3.7 Model fitting

All parametric methods employ the general principle of model fitting. The particular assumptions that they are based on can be regarded as a particular model. The model contains specific parameters, and the goal of the data analysis is to estimate these parameters. The parameters are estimated by fitting the model to the data.

EXAMPLE As a very simple example, consider a single sample of data. If it is assumed that these data follow a normal distribution, then the model is simply the normal distribution. The model contains two parameters, the mean and the variance. Depending on which values are chosen for the parameters, the agreement between the distribution and the data is better or worse. The question now is “what parameter values give the best agreement between the model and the data, i.e. give the best fit of the model to the data?”.

To be able to answer the latter question, we have to define a measure for the “distance” between the data and the model, to be used as the fit criteria. A very general criterion is the likelihood. This measure directly follows from the assumed distribution, and is applicable to whatever distribution is assumed. The likelihood criteria should be maximized, and when this is achieved, the associated parameter values are called maximum likelihood estimates.

Another much used fit criterion is the residual Sum of Squares (SS). This measure is defined as the sum of the squared residuals, i.e. the differences of each separate observed response with its associated expected response (according to the model). The best fit is found by minimizing the SS. In the simple example of fitting a normal distribution to a single dataset, the residuals are simply the differences of the observations to the mean. By changing the value of the mean, the SS varies.

The value of the mean resulting in the best fit, is exactly the value of the (arithmetic) sample mean. Put another way, the sample mean is the estimate of the mean of a normal distribution that results in the best fit according to the SS criteria. In the special case of a normal distribution, the sample mean is at the same time the maximum likelihood estimate. In other distributions however, the maximum likelihood or minimizing the SS results in different estimates of the parameters. For instance, for quantal dose-response data, the sum of squares is not appropriate, and the likelihood is the usual fit criteria.

---

4) Tied data are two or more observations of the same value. Parametric methods do exist for tied data, but these are beyond the scope of this Technical Specification.

The same principle of model fitting holds for more complicated models than a single dataset. For example, by replacing the mean of the normal distribution by a function of the dose, we obtain a dose-response model. Here, fitting the model by minimizing the SS or by maximizing the likelihood results in the same fit (because of the normality assumption).

A general method of finding the best fit is by trial and error, i.e. in an iterative search one tries to improve the likelihood by changing the parameter values, until an improvement can no longer be found. General algorithms exist that can perform such an iterative search in an efficient way. In particular models ("linear" models), the maximum likelihood estimates can be derived from explicit formulae and search algorithms are not required (for that reason linear models used to be popular before the availability of computers). In non-linear models, search algorithms can hardly be avoided. Although the user need not worry about the calculations underlying these algorithms, fitting non-linear models does require some basic understanding of the general principles of search algorithms (see 6.7).

4.3.8 Model checking

4.3.8.1 Analysis of residuals

After a model has been fitted to the data, a final check for the appropriateness of the fitted model may be performed. Do the data indeed comply with the model assumptions?

For instance,

— Do the data comply with the assumed distribution (in parametric analyses)?

— Are the variances homogenous (e.g. in ANOVA)?

— Is the dose-response model suitable for the dose-response data at hand (in dose-response analysis)?

A general approach for checking such assumptions is the analysis of residuals: the differences between the observations and the value predicted by model. For instance, in ANOVA the predicted value is the associated group mean, while in dose-response modelling, it is the value of the model at the relevant dose.

To check the distribution, the residuals can be taken together and be plotted in a single histogram, or in a (distribution-specific) QQ-plot[5)]. Visual inspection of such plots may reveal deviations from the assumed distribution, in particular when inspecting a QQ-plot, which should be linear if the data comply with the assumed distribution. Formal tests exist as well (see Clause 5), but it should be noted that a mild violation of the assumptions is no reason for concern, and tests do not measure the degree of violation.

Various other plots of the residuals can be made, e.g.

— against the predicted value (i.e. the group means, usually), to check if the variances are homogenous (if such were assumed);

— against the model prediction, to check for systematic deviations from the fitted model;

— against other experimental factors, if relevant, for instance the order in time by which the observations were made. Such a plot may show if the pertinent factor influenced the response systematically.

Finally, one may perform a formal goodness-of-fit test. This test is sensitive for all the assumptions simultaneously. A significant overall test of goodness-of-fit may indicate that one of the assumptions is not met, but this does not necessarily imply that the model is not useful for the particular purpose of the analysis. Here, one should judge the nature of the violated assumption and its potential impact on the results one is interested in. On the other hand, a nonsignificant goodness-of-fit test does not imply that the model used may be regarded as reliable. The test is more easily passed when the data contain relatively little information, and, as

5) The QQ-plot corresponds to a plot of observed quantiles versus expected quantiles.

a result, various models may pass the goodness-of-fit test, but lead to different conclusions. In other words, not only the model, but also the data should be "validated", by asking the question of whether or not they contain the information needed for answering the question of interest. The evaluation of a dose-response model for describing a particular dataset is more fully discussed in 6.4.

4.3.8.2 Validation of fitted dose-response model

In dose-response modelling, it may happen that the data appear to be unsuitable for that approach. This would happen if the dose-response information is too weak to have faith in any fitted dose-response model (see 6.4). For example, there may be large gaps between response levels or too few dose levels in areas where the response changes rapidly. Therefore, the estimation of an EC_x is only warranted if the dose-response data contain sufficient information on the shape of the dose-response relationship.

4.3.9 Reporting the results

The final step in a statistical analysis is reporting the results. Basically, two types of information should be given: the results of the analysis, and the justification of the methods (assumptions) used.

The results of the analysis typically consist of summarizing statistics. In current practice, these are usually the means and standard deviations (or standard errors) per dose group. This may not generally be the best way of reporting results, however.

When a parametric analysis assuming homogenous variances is applied, it is more informative to report the estimate of the common variance (residual Mean Square), together with a justification of the homogeneity assumption (e.g. a plot of the individual data or of the residuals against dose).

When a log-transformation is applied before the analysis, it is more adequate to report the geometric means, and the (possible common) geometric standard deviation (GSD) or Coefficient of Variation (CV).

When an NOEC is assessed, the associated test used should be reported, along with the test outcomes.

In the case where an EC_x is assessed, the fitted model should be reported, as well as the justification that the model was acceptable for assessing the EC_x (see 6.4).

More specific guidance for reporting results is given at the end of Clauses 4, 5 and 6.

5 Hypothesis testing

5.1 Introduction

5.1.1 General

This clause provides an overview of both hypothesis testing and methodological issues specific to determining NOECs under various experimental scenarios. It is divided into three major parts.

The first part includes flow-charts summarizing possible schemes for analysing quantal (Figure 3) and continuous data (Figures 4 and 5), along with some basic concepts that are important to the understanding of hypothesis testing and its use in the determination of NOECs. Special attention is given to the choice of the hypothesis to be tested, as this choice may vary depending on whether or not a simple dose-response trend is expected, and on whether increases, or decreases (or both) in response are of concern.

The remainder of the clause is divided into two major subclauses that discuss statistical issues related to the determination of NOECs for quantal and continuous data (4.2 and 4.3 respectively) and provide further details on the methods listed in Figures 3 and 4.

This division reflects the fact that different statistical methods are required for each type of data, and that problems arise that are unique to the analysis of each type of data. An attempt has been made to mention the

most widely used statistical methods, but to focus on a set of methods that combine desirable statistical properties with reasonable simplicity. For a given set of circumstances, more than one statistical approach may be acceptable. In such cases, the methods are described, the limitations and advantages of each are given, and the choice is left to the reader.

The flow-charts in Figures 3 and 4 indicate a possible choice of methods. Examples of the application of many of these methods, mathematical details and properties of the methods are presented in E.1.

The most commonly used methods for determining the NOEC are not necessarily the best. Relatively modest changes in current procedures for determining NOECs (e.g. selection of more powerful or biologically more plausible statistical methods) can improve the scientific basis for conclusions, and result in conclusions that are more protective of both the environment and business interests. Thus, some of the methods recommended may be unfamiliar to some readers, but all of the recommended methods should be compatible with current ISO and OECD guidelines that require the determination of NOECs.

A basic principle in selecting statistical methods is to attempt to use underlying statistical models that are consistent with the actual experimental design and underlying biology. This principle has historically been tempered by widely adopted conventions. For example, it is traditional in ecotoxicological studies to analyse the same response measured at different time points separately by time point, although in many cases unified analysis methods may be available. It is not the purpose of this subclause to explore this issue. Instead, discussion is restricted to the most appropriate analysis of a response at a single time point and, usually, for a single sex.

NOECs, as defined and discussed in this Technical Specification, are based on a concept sometimes called “proof of hazard”. In essence, the test substance is presumed to be non-toxic unless the data present sufficient evidence to conclude toxicity.

Alternative approaches to assessing toxicity through hypothesis testing exist. For example, Tamhane *et al.* (2001) and Hothorn and Hauschke (2000) developed an approach based on proof of non-hazard. Specifically, if an acceptable threshold of effect is specified, such as a 20 % decrease in mean, then the maximum safe dose (MAXSD in Tamhane *et al.,* 2001) is the highest concentration for which there is significant evidence that the mean effect is less than 20 %. These are relatively new approaches that have not been thoroughly tested in a practical setting, and for few endpoints is there agreement on what level of effect is biologically important to detect. All current guidelines regarding NOEC are based on the proof of hazard concept. For these reasons, this alternative approach is not presented in this clause, though it does hold some promise for the future. The only common exception to this is in regard to limit tests, where in addition to determining whether there is a statistically-significant effect in the single test concentration, one also tests for whether the effect in the test concentration is less than 50 %. A simple t-test can be used for that purpose.

It should also be realized that statistics and statistical significance cannot be solely viewed as representative of biological significance. There can be no argument that statistical significance (or lack thereof) depends on many factors in addition to the magnitude of effect at a given concentration. Statistics is a tool that is used to aid in the determination of what is biologically significant. If an observed effect is not statistically significant, the basis for deciding that it is nonetheless biologically significant is, obviously, not statistical. Lack of statistical significance may be because of a low power test. On the other hand, a judgment of biological significance without sufficient data to back it up is questionable.

The flow-charts and methodology presented indicate preliminary assessment of data to help guide the analysis. For example, assessments of normality, variance homogeneity, and dose-response monotonicity are advocated routinely. Such preliminary assessments do affect the power characteristics of the subsequent tests. The alternative to making these assessments is to ignore the characteristics of the data to be analysed. Such an approach can be motivated on the perceived general characteristics of each endpoint. However, this does not avoid the penalty of sometimes using a low power or inappropriate method when the data do not conform to expectation. A bias of this clause is to examine the data to be analysed and use this examination to guide the selection of the formal test to be applied. The preliminary assessment can be through formal tests or informed by expert judgment or some combination of the two. Certainly expert judgment should be employed whenever feasible, and when used, is invaluable to sound statistical analysis. These charts provide guidance, but sound statistical judgment sometimes leads to departures from the flow-charts.

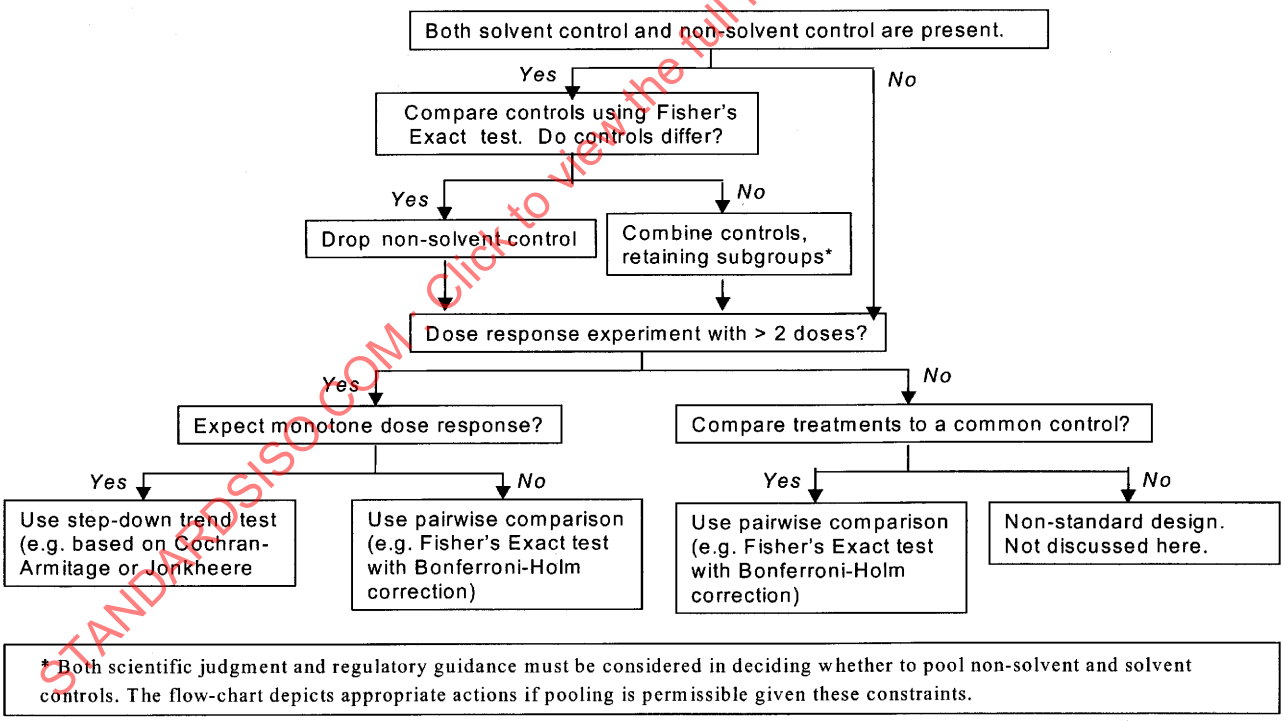
The flow-charts (Figures 3 and 4) are intended to include experiments which contain only two concentrations (control and one test concentration). Such experiments are generally referred to as limit tests and the methods described are applicable to these tests.

It should be noted that tests of hypotheses might also be required for various special-case assessments of study results (e.g. use of a contingency table to assess the significance of male-female differences in frequency of responses at some dose). These types of analyses are beyond the scope of this Technical Specification.

The terms “dose” and “concentration” are used interchangeably in this clause and the control is a zero dose or zero concentration group. Consistent with this, the terms “doses” and “concentrations” include the control, so that, for example, an experiment with only two concentrations has one control group and one positive concentration group.

The tests discussed in this clause, with the exception of the Tamhane-Dunnett and Dunn tests, are all available in commercial software. For example, they are available in SAS version 8 and higher[6)]. The two-sided Tamhane-Dunnett test (though not called as such) is available in SAS through the studentized maximum modulus distribution provided by the probmc function. Where these tests are discussed, alternatives are provided, so that the reader can follow the general guidance of this clause without being forced to develop special programs.

It is observed that there is no special flow-chart for the exact Jonckheere-Terpstra and exact Wilcoxon tests. One of the appealing features of these two tests is that there are both asymptotic and exact versions and the same logic applies to both.



NOTE Note that the dose count in ‘> 2’ includes the control.

Figure 3 — Analysis of quantal data: Methods for determining the NOEC

6) SAS version 8 is an example of a suitable product available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of this product.

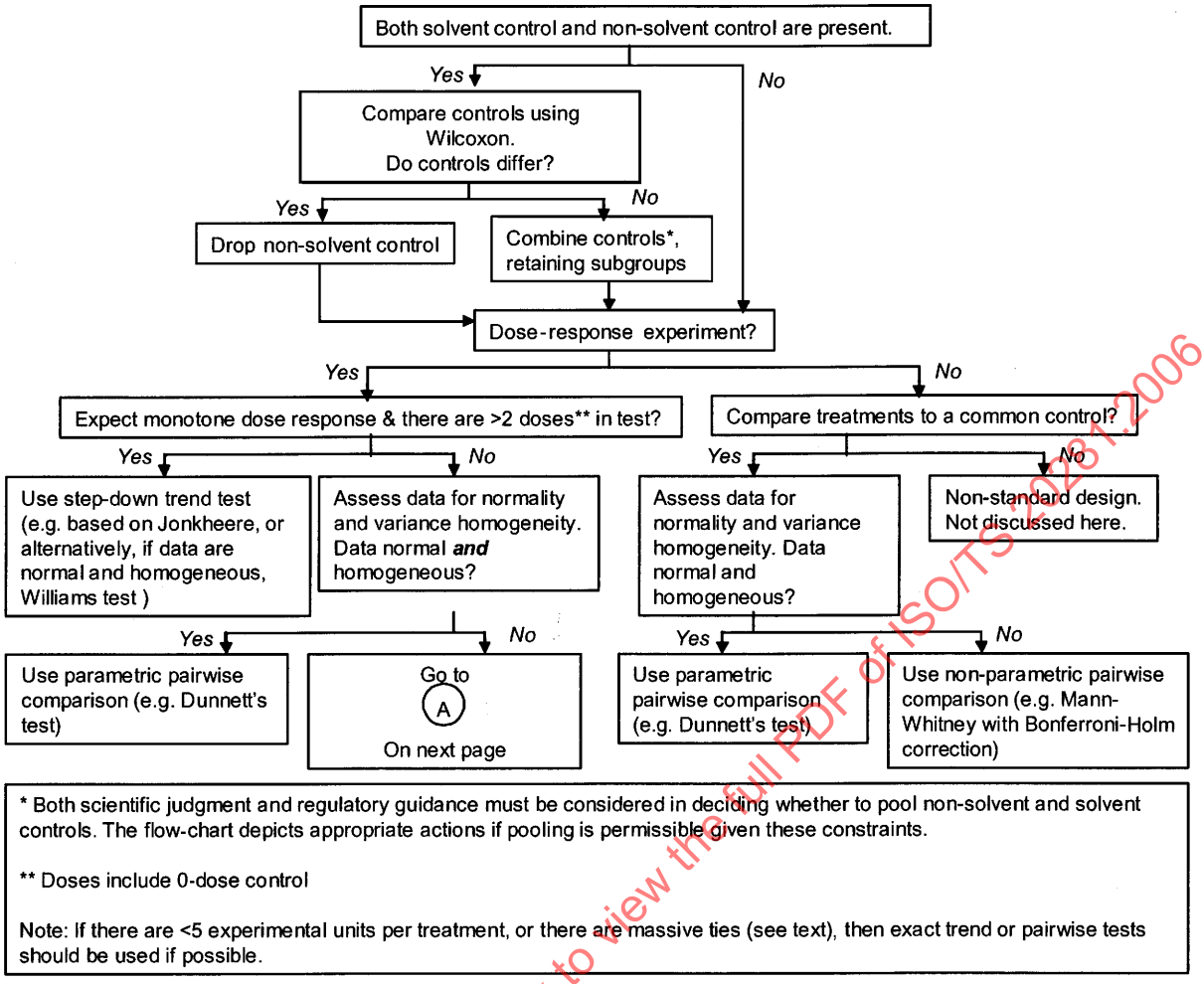


Figure 4 — Analysis of continuous data: Methods for determining the NOEC

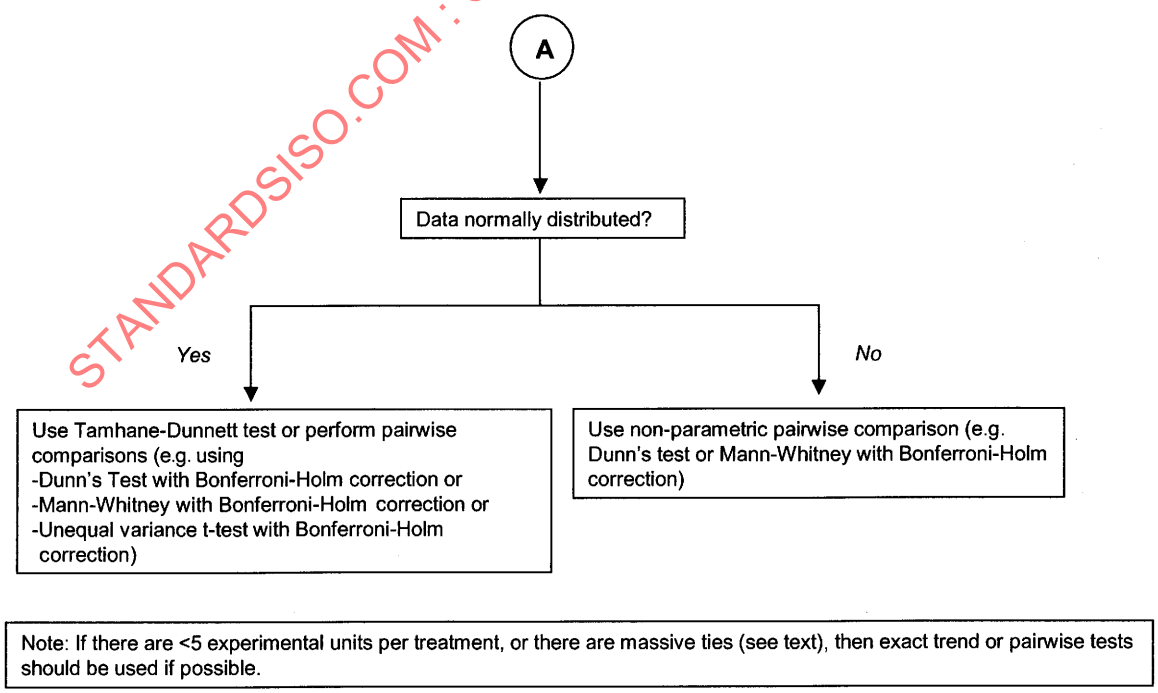


Figure 5 — Analysis of continuous data: Methods for determining the NOEC (*continued*)

5.1.2 NOEC: What it is, and what it is not

The NOEC (3.18) is defined as the test concentration below the lowest concentration that resulted in a significant effect in the specific experiment, i.e. the NOEC is the tested concentration just below the LOEC.

A significant effect is generally meant to be a statistically significant effect, as resulting from a hypothesis test. Obviously, no claim can be made that the condition of organisms exposed to toxicants at the NOEC is the same as the condition of organisms in the control group, or that the NOEC is an estimate of the threshold of toxicity (if such exists). Rather, no effect could be detected in this particular experiment. The detectability of an effect depends on the quality and the size of the experiment and the statistical procedure used. Of course, zero effects are never detectable. The relationship between the detectability of effects and the quality of the experiment can be quantified by the concept of statistical power. For a given null and alternative hypothesis, sample size and variance, statistical power is the probability that a particular magnitude of effect results in a significant test outcome. In large experiments (i.e. many replicates), smaller sized effects are detectable as compared to small experiments. Thus, one may consider the detectable effect size of a particular experiment as an analogue of the detection limit of a particular chemical analysis. The detectable effect size can be increased not only by using larger sample sizes, but also by taking measures to make the experimental (residual) error smaller and by selecting more powerful statistical tests.

Power calculations are useful for the purpose of designing experiments in such a way that effect sizes that are considered relevant are likely to be (statistically) detected. Care shall be taken when using information on the power for interpreting an NOEC. If the test was designed to detect a difference of x %, and an observed treatment effect is not found to be statistically significant, this does not allow one to conclude with a specified level of confidence that the true effect in the population is less than x %.

Meaningful confidence intervals for the effect size at a given concentration are sometimes possible. An application of this is discussed in 4.1.4 and methods for doing this are developed in E.3. For some techniques, obtaining meaningful confidence intervals is very difficult and this is discussed in greater detail in that annex.

5.1.3 Hypothesis used to determine NOEC

5.1.3.1 Understanding the question to be answered

The hypothesis that is tested in determining the NOEC for a toxicological experiment reflects the risk assessment question and the assumptions that are made concerning the underlying characteristics, or statistical model, of the responses being analysed (e.g. does the response increase in an orderly, i.e. monotone way with increasing toxicant concentration?).

The statistical test that is used depends on

— the hypothesis tested (e.g. are responses in all the groups equal?),

— the associated statistical model, and

— the distribution of the values (e.g. are data normally distributed?).

Thus, it is necessary to understand the question to be answered and to translate this question into appropriate null and alternative hypotheses before selecting the test procedure.

The need to select a statistical model for assessing the results of toxicity tests is not unique to the hypothesis testing approach. All methods of assessment assume a statistical model. The hypothesis testing approach to the evaluation of toxicity data is based, in part, on keeping to a reasonable number the untestable or difficult-to-test assumptions, particularly those regarding the statistical model that is used in reaching conclusions. The models used in regression and biology-based methods use stronger assumptions than the models used in the hypothesis-testing approach.

The simplest statistical model generally used in hypothesis testing assumes only that the distributions of responses within these populations are identical except for a location parameter (e.g. the mean or median of the distribution of values from each group).

Another statistical model that is often used assumes that there is a trend in the response that is associated with increasing exposure.

Each of these models suggests a set of hypotheses that can be tested to determine whether the model is consistent with the data. These two types of hypotheses can further be expressed as one-sided or two-sided. The discussion below is developed in terms of population means, but applies equally to hypotheses concerning population medians.

5.1.3.2 One-sided hypothesis

The most basic hypothesis (in one-sided form) can be stated as follows:

$H_0 : \mu_0 = \mu_1 = \mu_2 = \ldots = \mu_k$ versus $H_1 : \mu_0 > \mu_i$ for at least one i, (Model 1)

where

μ_i, $i = 0, 1, 2, 3, \ldots, k$ denote the means of the control and test *populations*, respectively.

Thus, one tests the null hypothesis of no differences among the population means against the alternative that at least one population mean is smaller than the control mean. There is no investigation of differences among the treatment means, only whether treatment means differ from the control mean. The one-sided hypothesis is appropriate when an effect in only one direction is a concern. The direction of the inequality in the above alternative hypothesis (i.e. in $H_1 : \mu_0 > \mu_i$) would be appropriate if a decrease in the endpoint was a concern, but an increase was not (for instance, if an exposure was expected to induce infertility and reduce the number of offspring). If an increase in the endpoint was the only concern, then the direction of the inequality would be reversed.

5.1.3.3 Two-sided trend test

In the two-sided form of the hypothesis, the alternative hypothesis is:

$H_1 : \mu_0 \neq \mu_i$ for at least one i.

5.1.3.4 Trend or pair-wise test

If no assumption is made about the relationships among the treatment groups and control (e.g. no trend is assumed), the test statistics are based on comparing each treatment to the control, independent of the other treatments. Many tests have been developed for this approach, some of which are discussed below. Most such tests were developed for experiments in which treatments are qualitatively different, as, for example, in comparing various new therapies or drug formulations to a standard.

In toxicology, the treatment groups generally differ only in the exposure concentration (or dose) of a single chemical. It is further often true that biology suggests that if the chemical is toxic, then as the level of exposure is increased, the magnitude effect tends to increase. Depending on what response is measured, the effect of increasing exposure may show up as an increase or as a decrease in the measured response, but not both.

The statistical model underlying this biological expectation is what is called a trend model or a model assuming monotonicity of the population means:

$\mu_0 \geqslant \mu_1 \geqslant \mu_2 \geqslant \mu_3 \geqslant \ldots \geqslant \mu_k$ (or with inequalities reversed) (Model 2)

The null and alternative hypotheses can then be stated as

$$H_{02} : \mu_0 = \mu_1 = \mu_2 = \ldots = \mu_k \quad \text{versus} \quad H_{12} : \mu_0 \geqslant \mu_1 \geqslant \mu_2 \geqslant \mu_3 \geqslant \ldots \geqslant \mu_k \text{ , with } \mu_0 > \mu_k \text{ .}$$

Note that $\mu_0 \geqslant \mu_k$ is equivalent, under the alternative, to $\mu_0 > \mu_i$ for at least one i. If this monotone model is accepted as representing the true responses of test organisms to exposure to toxicants, it is not possible for, say, μ_3 to be smaller than μ_0 , and μ_6 not to be smaller.

Under the trend model and tests designed for that model, if tests of hypotheses H_{02} vs. H_{12} reveal that μ_3 is different from μ_0, but μ_2 is not, the NOEC has been determined (i.e. it is the test concentration associated with μ_2), and there is no need to test whether μ_1 differs from μ_0. Also, finding that μ_3 differs from μ_0 implies that a significant trend exists across the span of doses including μ_0 and μ_3, the span including μ_0 and μ_4, and so on.

For the majority of toxicological studies, a test of the trend hypothesis based on Model 2 is consistent with the basic expectations for a model for dose response. In addition, statistical tests for trend tend to be more powerful than alternative non-trend tests, and should be the preferred tests if they are applicable. Thus, a necessary early step in the analysis of results from a study is to consider each endpoint, decide whether a trend model is appropriate, and then choose the initial statistical test based on that decision. Only after it is concluded that trend is not appropriate, do specific pair-wise comparisons make sense to illuminate sources of variability.

Toxicologists sometimes do not know whether a compound causes measurements of continuous variables such as growth or mass to increase or decrease, but they are confident it acts in only one direction. For such endpoints, the two-sided trend test is appropriate, described in 4.1.7. One difference between implementing step-down procedures for quantal data and continuous data is that two-sided tests are much more likely to be of interest for continuous variables. Such a model is rarely appropriate for quantal data, as only increased incidence rate above background (control) incidence are of interest in toxicology.

The two-sided version of the step-down procedure is based on the underlying model:

$$\mu_0 \geqslant \mu_1 \geqslant \mu_2 \geqslant \mu_3 \geqslant \ \ldots \ \geqslant \mu_k$$

or

$$\mu_0 \leqslant \mu_1 \leqslant \mu_2 \leqslant \mu_3 \leqslant \ldots \leqslant \mu_k.$$

Under this model, in testing the hypothesis that all population means are equal against the alternative (that at least one inequality is strict), one first tests separately each one-sided alternative at the 0,025-level of significance with all doses present.

— If neither of these tests is significant, the NOEC is higher than the highest concentration. If both of these tests are significant, a trend-based procedure should not be used, as the direction of the trend is unclear.

— If exactly one of these tests with all the data is significant, then the direction of all further tests is in the direction of the significant test with all groups.

Thereafter, the procedure is as in the one-sided test, except all tests are at the 0,025 significance level to maintain the overall 0,05 false positive rate.

Where it is biologically sensible, it is preferable to test the one-sided hypothesis, because random variation in one direction can be ignored, and as a result, statistical tests of the one-sided hypothesis are more powerful than tests of the two-sided hypothesis.

Note that a hypothesis test based on Model 2 assumes only a monotone dose response rather than a precise mathematical form, such as is required for regression methods (Clause 6) or the biology-based models (Clause 7).

5.1.4 Comparisons of single-step (pair-wise comparisons) or step-down trend tests to determine the NOEC

5.1.4.1 General discussion

In general, determining the NOEC for a study involves multiple tests of hypotheses (i.e. a family of hypotheses is tested), and either pair-wise comparisons of treatment groups, or a sequence of tests of the significance of trend. For these reasons, statisticians have developed tests to control the family-wise error rate, FWE (the probability that one or more of the null hypotheses in the family are rejected incorrectly), in the multiple comparisons performed to identify the NOEC.

EXAMPLE For example, suppose one compares each of ten treatments to a common control using a simple t-test with a false positive error rate of 5 % for each comparison. Suppose further that none of the treatments has an effect, i.e. all of the treatment and control population means are equal. For each comparison, there is a 5 % chance of finding a significant difference between that sample treatment mean and the control. The chance that at least one of the ten comparisons is wrongly declared significant is much higher, possibly as high as $1 - 0{,}95^{10} = 0{,}4$ or 40 %.

The method of controlling the family-wise error rate has important implications for the power of the test. There are two approaches that are discussed: single-step procedures and step-down procedures. There are numerous variations within each of these two classes of procedures that are suited for specific data types, experimental designs and data distributions.

A factor that shall be considered in selecting the methods for analysing the results from a study is whether the study is a dose-response experiment. In this context, a dose-response experiment is one in which treatments consist of a series of increasing doses of the same test material. Monotone responses from a dose-response experiment are best analysed using step-down procedures based on trend tests (e.g. the Cochran-Armitage, Williams, or Jonckheere-Terpstra trend test), whereas non-monotone responses shall be analysed by pair-wise comparisons to the control (e.g. Fisher's Exact test or Dunnett's test). This subclause discusses when to use each of these two approaches.

5.1.4.2 Single-step procedures

Single-step procedures amount to performing all possible comparisons of treatment groups to the control. Multiple comparisons to the control may be made, but there is no ordered set of hypotheses to test, and no use of the sequence of outcomes in deciding which comparisons to make. Examples of the single-step approach include the use of the Fisher's Exact test, the Mann-Whitney, Dunnett and Dunn tests. Since many comparisons to the control are made, some adjustment shall be made for the number of such comparisons to keep the family-wise error (FWE) rate at a fixed level, generally 0,05. With tests that are inherently single comparison tests, such as Fisher's exact and Mann-Whitney, a Bonferroni adjustment can be made: a study with k treatment levels would be analysed by performing the pair-wise comparisons of each of the treatment groups to the control group, each performed at a significance level of α/k instead of α. (This is the Bonferroni adjustment.) Equivalently, the calculated p-value ignoring multiplicities is multiplied by k. That is, $p^b_i = k^*p_i$. The Bonferroni adjustment is generally overly conservative, especially for large k. Modifications reduce the conservatism while preserving the FWE at 0,05 or less.

For the Holm modification of the Bonferroni adjustment, arrange the k unadjusted p-values for all comparisons of treatments to control in rank order, i.e. $p_{(1)} \leqslant p_{(2)} \leqslant p_{(3)} \leqslant \ldots \leqslant p_{(k)}$. Beginning with $p_{(1)}$, compare $p_{(i)}$ with $\alpha/(k - I + 1)$, stopping at the first non-significant comparison. If the smallest i for which $p_{(i)}$ exceeds $\alpha/(k - I + 1)$ is $I = j$, then all comparisons with $I > j$ are judged non-significant without further comparisons. It is helpful (Wright, 1992) to report adjusted p-values rather than the above comparisons. Thus, report $p^*_{(1)} = p_{(1)}{}^*(k - I + 1)$ and then compare each adjusted p-value to α. Table 2 illustrates the advantage of the Bonferroni-Holm method. In this hypothetical example, only the comparison of Treatment 4 with the control would be significant if the Bonferroni adjustment is used, whereas all comparisons except the comparison of the Control with Treatment 1 would be significant if the Bonferroni-Holm adjustment is used.

Table 2 — Comparison of adjusted and unadjusted p-values

Comparison	Unadjusted p value	Bonferroni-Holm Adjusted p value $p^{*}_{(i)}$	Bonferroni Adjusted p-values $P^{b}{}_{i}$
Control – Treatment 4	$p_{(1)} = 0{,}002$	$0{,}002^{*}4 = 0{,}008$	$0{,}002^{*}4 = 0{,}008$
Control – Treatment 2	$p_{(2)} = 0{,}013$	$0{,}013^{*}3 = 0{,}039$	$0{,}013^{*}4 = 0{,}052$
Control – Treatment 3	$p_{(3)} = 0{,}020$	$0{,}020^{*}2 = 0{,}040$	$0{,}02^{*}4 = 0{,}08$
Control – Treatment 1	$p_{(4)} = 0{,}310$	$0{,}310*1 = 0{,}310$	$0{,}310^{*}4 = 1$

Alternatives based on the Sidak inequality [each comparison at level $1 - (1 - \alpha)^{k}$] are also available. The Bonferroni and Bonferroni-Holm adjustment guarantee that the family-wise error rate is less than α, but they are conservative. Other tests, such as Dunnett's, have a "built-in" adjustment for the number of comparisons made and are less conservative (hence, more powerful). For completeness, it should be understood that if only one comparison is made, the Bonferroni and Bonferroni-Holm adjustments leave the p-value unchanged. Of course, there is no need to refer to an adjustment in this simple case, but the discussion becomes needlessly complicated if special reference is always made to the case of only one comparison.

5.1.4.3 Step-down procedures

Step-down procedures are generally preferred where they are applicable. All step-down procedures discussed are based on a sequential process consisting of testing an ordered set of hypotheses concerning means, ranks, or trend. A step-down procedure based on trend (for example) works as follows.

— First, the hypothesis that there is no trend in response with increasing dose is tested when the control and all dose groups are included in the test.

— Then, if the test for trend is significant, the high dose group is dropped from the data set, and the hypothesis that there is no trend in the reduced data set is tested.

This process of dropping treatment groups and testing is continued until the first time the trend test is non-significant. The highest dose in the reduced data set at that stage is then declared to be the NOEC. Distinguishing features of step-down procedures are that the tests of hypothesis shall be performed in a given order, and that the outcome of each hypothesis test is evaluated before deciding whether to test the next hypothesis in the ordered sequence of hypotheses. It is these two aspects of these procedures that account for controlling the family-wise error (FWE) rate.

A step-down method typically uses a critical level larger than that used in single-step procedures, and seeks to limit the number of comparisons that need to be made. Indeed, the special class of "fixed-sequence" tests described below fix the critical level at 0,05 for each comparison but bound the FWE rate at 0,05. Thus, step-down methods are generally preferable to the single-step methods as long as the response means are monotonic.

Tests based on trend are logically consistent with the anticipated monotone pattern of responses in toxicity tests. Step-down procedures make use of this ordered alternative by ordering the tests of hypotheses. This minimizes the number of comparisons that need to be made, and in all the methods discussed here, a trend model is explicitly assumed (and tested) as a part of the procedure.

Procedures that employ step-down trend tests have more power than procedures that rely on multiple pair-wise comparisons when there is a monotone dose response because they make more use of the biology and experimental design being analysed. When there is a monotone dose response, procedures that compare single treatment means or medians against the control, independent of the results in other treatments (i.e. single-step procedures), ignore important and relevant information, and suffer power loss as a result.

The trend models used in the step-down procedures do not assume a particular precise mathematical relationship between dose and response, but rather use only monotonicity of the dose-response relationship.

The underlying statistical model assumes a monotone dose response in the *population* means, not the *observed* means.

Rejection of the null hypothesis (i.e. rejecting the hypothesis that all group means, or medians, or distributions are equal) in favour of the stated alternative implies that the high dose is significantly different from the control. The same logic applies at each stage in the step-down application of the test to imply, whenever the test is significant, that the high dose remaining at that stage is significantly different from the control. These tests are all applied in a one-sided manner with the direction of the alternative hypothesis always the same. Moreover, this methodology is general, and applies to any legitimate test of the stated hypotheses under the stated model. That is, one can use this fixed-sequence approach with the Cochran-Armitage test on quantal data, and the Jonckheere-Terpstra or Williams or Brown-Forsythe tests of trend on continuous data. Other tests of trend can also be used in this manner.

5.1.4.4 Deciding between the two approaches

Bauer (1997) has shown that certain tests based on a monotone dose response can have poor power properties or error rates when the monotone assumption is wrong. For example, departures from monotonicity in non-target plant data are common, where they arise from low dose stimulation. Davis and Svendsgaard (1990) suggest that departures from monotonicity may be more common than previously thought. These results suggest that a need for caution exists. There are two testing philosophies used to determine whether a monotone dose response is appropriate. Some recommend assessing in a general way for an endpoint or class of endpoints, whether a monotone dose response is to be expected biologically. If a monotone trend is expected, then trend methods are used. This procedure should be augmented, at a minimum, by adding that, if a cursory examination of the data shows strong evidence of departure from monotonicity (i.e. large, consistent departures), then pair-wise methods should be used instead.

A second philosophy recommends formal tests to determine if there is significant monotonicity or significant departure from monotonicity. With continuous data, one can use either

a) a positive test for monotonicity (such as Bartholomew's test) and proceed only if there is evidence of monotonicity, or

b) use a "negative" test for departure from monotonicity (such as sets of orthogonal contrasts for continuous responses and a decomposition of the Chi-squared test of independence for quantal responses) and proceed unless there is evidence of non-monotonicity.

Details on these procedures are given in E.1 and E.3. Either philosophy is acceptable. The second approach is grounded in the idea that monotonicity is the rule, and that it should take strong evidence to depart from this rule. Both approaches reduce the likelihood of having to explain a significant effect at a low or intermediate concentration when higher concentrations show no such effect. The "negative" testing approach is more consistent with the way tests for normality and variance homogeneity are used and is more likely to result in a trend test than a method that requires a significant trend test to proceed. This is what is shown in the flow diagrams presented below.

Formal tests for monotonicity are especially desirable in a highly automated test environment.

One simple procedure that can be used in this situation for continuous responses is to construct linear and quadratic contrasts of normalized rank statistics (to avoid the complications that can arise from non-normal or heterogeneous data). If the linear contrast is not significant and the quadratic contrast is significant, there is evidence of possible non-monotonicity that calls for closer examination of the data or pair-wise comparison methods. Otherwise, a trend-based analysis is used.

A less simple, but more elegant procedure would be to construct simultaneous confidence intervals for the mean responses assuming monotonicity (i.e. isotonic estimators based on maximum likelihood criteria; see E.3) and use a trend approach unless one or more sample (i.e. non-isotonic) means fall outside the associated confidence interval. For quantal data using the Cochran-Armitage test, there is a built-in test for lack of monotonicity.

Where expert judgement is used, formal tests for monotonicity or its lack may be replaced by visual inspection of the data, especially of the mean or median responses. The same concept applies to assessing normality and variance homogeneity.

5.1.5 Dose metric in trend tests

Various authors have evaluated the influence on trend tests of the different ways of expressing dose (i.e. dose metrics), including actual dose-values, log(dose), and equally-spaced scores (i.e. rank-order of doses). Lagakos and Lewis (1985) discuss various dose metrics and prefer the rank-order as a general rule. Weller and Ryan (1998) likewise prefer rank ordering of doses for some trend tests.

When dose values are approximately equally spaced on a log scale, there is little difference between using log(dose) and rank-order, but use of actual dose values can have the unintended effect of turning a trend test into a comparison of high dose to control, eliminating the value of the trend approach and compromising its power properties. This is not an issue with some tests, such as the Jonckheere-Terpstra test discussed below, since rank-order of treatment groups is built into the procedure. With others, such as the Cochran-Armitage and contrast-based tests, it is an important consideration.

Extensive computer simulations have been done (J.W. Green, in preparation) to compare the use of rank-order to dose-value in the Cochran-Armitage test. One simulation study involved over 88 000 sets of dose-response scenarios for 4- and 5-dose experiments found 12 % to 17 % of the experiments where the rank-order scoring found lower NOEC than dose-value did, and only 1 % of the experiments where dose-value scores led to a lower NOEC than when the rank-order scores were used. In the remaining cases, the two methods established the same NOEC. While these simulations results do not, by themselves, justify the use of rank-order over actual dose levels or their logarithms, they do suggest that use of rank-order does not lessen the power of statistical tests. All trend-based tests discussed in this Technical Specification, including contrast tests for monotonicity, are based on rank ordering of doses.

5.1.6 Role of power in toxicity experiments

The adequacy of an experimental design and the statistical test used to analyse study results are often evaluated in terms of the power of the statistical test. Power is defined as the probability that a false null hypothesis is rejected by the statistical test in favour of a true alternative. That power depends on the alternative hypothesis. In the context of toxicology, the larger the effect, the higher the power to detect that effect. So, if a toxicant has had some effect on the organisms in a toxicity test, power is the probability that a difference between treatment groups and the control is detected. The power of a test can be calculated if we know the size of the effect to be detected, the variability of the endpoint measured, the number of treatment groups, and the number of replicates in each treatment group. (Detailed discussions are given in 5.2 and 5.3 and E.1 and E.3).

It should be understood that the goal of selecting a method for determining an NOEC is not to find the most powerful method. Rather, the focus should be on selecting methods most appropriate for the data and end result. Power is certainly an ingredient in this selection process. As discussed below, power can be used in designing experiments and selecting statistical tests to reduce animal use without loss of statistical power. This can be accomplished by selecting an inherently more powerful test applied to fewer animals, so that the result is to retain the power of more traditional tests but use fewer animals.

The primary use of power analysis in toxicity studies is in the design stage. By demonstrating that a study design and test method have adequate power to detect effects that are large enough to be deemed important, if we then find that, at a given dose, there is no statistically significant effect, we can have some confidence that there is no effect of concern at that dose. However, power does not quantify this confidence. Failure to adequately design or control an experiment so that statistical tests have adequate power can result in large effects being found to be statistically insignificant. On the other hand, it is also true that a test can be so powerful that it finds statistically significant effects of little importance.

Deciding on what effect size should be considered to be large enough to be important is difficult, and may depend on both biological and regulatory factors. In some cases, the effect size may be selected by regulatory agencies or specified in guidelines.

A requirement to demonstrate an adequate power to detect effects of importance removes any perceived reward for poor experimental design or technique, as poor experimental design is shown to have low power to detect important effects, and leads to the selection of more powerful statistical tests and better designs. The latter is preferable to the alternative of increasing sample sizes. Indeed, it is sometimes possible to find statistical procedures with greater power to detect important differences or provide improved estimates and simultaneously decrease sample sizes.

For design purposes, the background variance can be taken to be the pooled within-experiment variance from a moving frame of reference from a sufficiently long period of historical control data with the same species and experimental conditions. The time-window covered by the moving frame of reference should be long enough to average out noise without being so long that undetected experimental drift is reflected in the current average. If available, a three-to-five year moving frame of reference might be appropriate. When experiments shall be designed using more limited information on variance, it may be prudent to assume a slightly higher value than what has been observed. Power calculations used in designs for quantal endpoints shall take the expected background incidence rate into account for the given endpoint, as both the Fisher Exact and Cochran-Armitage tests are sensitive to this background rate, with highest power achieved for a zero background incidence rate. The background incidence rate can be taken to be the incidence rate in the same moving frame of reference already mentioned.

While at the design stage, power shall, of necessity, be based on historical control data for initial variance estimates, it may also be worthwhile to do a post-hoc power analysis as well to determine whether the actual experiment is consistent with the criteria used at the design stage. Care shall be taken in evaluating post-hoc power against design power. Experiment-to-experiment variation is expected and variance estimates are more variable than means. The power determination based on historical control data for the species and endpoint being studied should be reported.

Alternatively, for experimental designs constructed to give an acceptable power based on an assumed variance rather than on historical control data, a post-hoc test can be done to compare the observed variance to the variance used in designing the experiment. If this test finds significantly higher observed variance (e.g. based on a Chi-squared or F-test) than that used in planning, then the assumptions made at design time may need to be reassessed.

5.1.7 Experimental design

Factors that shall be considered when developing experimental designs include the number and spacing of doses or exposure levels, the number of subjects per dose group, and the nature and number of subgroups within dose groups. Decisions concerning these factors are made so as to provide adequate power to detect effects that are of a magnitude deemed biologically important.

The choice of test substance concentrations is one aspect of experimental design that shall be evaluated for each individual study. The goal is to bracket the NOEC with concentrations that are as closely spaced as practical. If limited information on the toxicity of a test material is available, test concentrations or doses can be selected to cover a range somewhat greater than the range of exposure levels expected to be encountered in the field and should include at least one concentration expected not to have a biologically important effect. If more information is available this range may be reduced, so that doses can be more closely spaced. Where effects are expected to increase approximately in proportion to the log of concentration, concentrations should be approximately equally spaced on a log scale. Three to seven concentrations plus concomitant controls are suggested, with the smaller experiment size typical for acute tests and larger experiment sizes most appropriate when preliminary dose finding information is skimpy.

The trade-off between number of subjects per subgroup and number of subgroups per group should be based on power calculations using historical control data to estimate the relative magnitude of within- and among-subgroup variation and correlation. If there are no subgroups, then there is no way to distinguish housing effects from concentration effects and neither between- and within-group variances nor correlations can be estimated, nor is it possible to apply any of the statistical tests described for continuous responses to subgroup means other than the Jonckheere-Terpstra test. Thus, a minimum of two subgroups per concentration is recommended; three subgroups are much better than two; four subgroups are better than three. The improvement in modelling falls off substantially as the number of subgroups increases beyond four. (This can be understood on the following grounds. The modelling is improved if we get better estimates of

both among- and within-subgroup variances. The quality of a variance estimate improves as the number of observations on which it is based increases. Either sample variance has, at least approximately, a Chi-squared distribution. The quality of a variance estimate can be measured by the width of its confidence interval and a look at a Chi-squared table verifies the statements made.) The precise needs for a given experiment depend on factors such as the relative and absolute size of the between- and within-replicate variances. Examples 1 and 2 in Annex E.3 illustrate the trade-offs between replicates per concentration and subjects per replicate.

In any event, the number of subgroups per concentration and subjects per subgroup should be chosen to provide adequate power to detect an effect of magnitude judged important to detect. This power determination should be based on historical control data for the species and endpoint being studied.

Since the control group is used in every comparison of treatment to control, consideration should be given to allocating more subjects to the control group than to the treatment groups in order to optimize power for a given total number of subjects. The optimum allocation depends on the statistical test to be used. A widely used allocation rule was given by Dunnett (1955), which states that for a total of N subjects and k treatments to be compared to a common control, if the same number, n, of subjects are allocated to every treatment group, then the number, n_0, to allocate to the control to optimize power is determined by the so-called square-root rule. By this rule, the value of n is (the integer part of) the solution of the equation $N = kn + n\sqrt{k}$, and $n_0 = \quad - kn$. (It is almost equivalent to say $n_0 = n\sqrt{k}$.) This has been shown to optimize power for Dunnett's test. It is used, often without formal justification, for other pair-wise tests, such as the Mann-Whitney and Fisher exact tests. Williams (1972) showed that the square-root rule may be somewhat sub-optimal for his test and optimum power is achieved when $\sqrt{k}$ in the above equation is replaced by something between $1{,}1\sqrt{k}$ and $1{,}4\sqrt{k}$.

The optimality of the square-root rule to other tests, such as Jonckheere-Terpstra and Cochran-Armitage has not been published in definitive form, but simulations (manuscript in preparation by J.W. Green) show that for the step-down Jonckheere-Terpstra test, power gains of up to 25 % are common under this rule compared to results from equal sample sizes. In all the cases examined, the power is greater following this rule compared to equal sample sizes, where the total sample size is held constant. In the absence of definitive information on the Jonckheere-Terpstra and other tests, it is probably prudent to follow the square-root rule for pair-wise, Jonckheere-Terpstra and Cochran-Armitage tests and either that or Williams' modification of the rule for other step-down procedures.

The selection of an allocation rule is further complicated in experiments where two controls are used, since if the controls are combined for further testing, a doubling of the control sample size is already achieved. Since experience suggests that most experiments find no significant difference between the two controls, the optimum strategy for allocating subjects is not necessarily immediately clear. This of course would not apply if a practice of pooling of controls is not followed.

The reported power increases from allocating subjects to the control group according to the square-root rule do not consider the effect of any increase in variance as concentration increases. One alternative, not without consequences in terms of resources and treatment of animals, is to add additional subjects to the control group without subtracting from treatment groups. There are practical reasons for considering this, since a study is much more likely to be considered invalid when there is loss of information in the controls than in treatment groups.

5.1.8 Treatment of covariates and other adjustments to analysis

It is sometimes necessary to adjust the analysis of toxicity data by taking into account some restriction on randomization, compartmentalization (housing) or by taking into account one or more covariates that might affect the conclusions.

EXAMPLE Examples of potential covariates include: initial body masses, initial plant heights and age at start of test.

While a thorough treatment of this topic is not presented, some attention to this topic is in order.

For continuous, normally distributed responses with homogeneous variances, analysis of covariance (ANCOVA) is well developed. Hocking (1985) and Milliken and Johnson (1984) are among the many

references on this topic. For continuous responses that do not meet the normality or homogeneity requirements, non-parametric ANCOVA is available.

Shirley (1981) indicates why non-parametric methods are needed in some situations. Stephenson and Jacobson (1988) contain a review of papers on the subject up to 1988. Subsequent papers include Wilcox (1991) and Knoke (1991). Stephenson and Jacobson recommend a procedure that replaces the dependent variable with ranks but retains the actual values of the independent variable(s). This has proved useful in toxicity studies. Seaman *et al*. (1985) discuss power characteristics of some non-parametric ANCOVA procedures.

When the response variable is quantal and is assumed to follow the binomial distribution, ANCOVA can be accomplished through logistic regression techniques. In this case, the covariate is a continuous regressor variable and the dose groups are coded as “dummy variables”. This approach can be more generally described in the Generalized Linear Model (GLM) framework (McCullagh and Nelder 1989). For quantal data, Koch *et al*. (1998), Thall and Vail (1990), Harwell and Serlin (1988), Tangen and Koch (1999a, 1999b) consider some relevant issues.

Adjustments shall be made to statistical methods when there are restrictions on randomization of subjects such as housing of subjects together. This is discussed for both quantal and continuous data in 5.2.2.5, 5.2.3, and 5.3.7.1, where the possibility of correlations among subjects housed together is considered, as are strategies for handling this problem. In the simple dose-response designs being discussed in this clause, other types of restrictions on randomization are less common. However, there is a large body of literature on the treatment of blocking and other issues that can be consulted. Hocking (1985) and Milliken and Johnson (1984) contain discussions and additional references.

Transformation of the doses (i.e. *not* response measures) in hypothesis testing is restricted, in this clause, to the use of rank order of the doses. For many tests, the way that dose values (actual or rank order) are expressed has no effect on the results of analysis. An exception is the Cochran-Armitage test. (See E.1.)

5.2 Quantal data (e.g. mortality, survival)

5.2.1 Hypothesis testing with quantal data to determine NOEC values

In this clause, the selection of methods and experimental designs for determining NOEC values focuses on identifying the tests most appropriate for detecting effects. The appropriateness of a given method hinges on the design of the experiment and the pattern of responses of the experimental units. Figure 3 illustrates an appropriate scheme for method selection, and identifies several statistical methods that are described in detail below. There are, of course, other statistical procedures that might be chosen. The following discussion identifies many of the procedures that might be used, gives details of some of the most appropriate, and attempts to provide some insight into the strengths and weaknesses of each method.

If there are two negative controls (i.e. solvent and non-solvent), Fisher’s Exact test, applied just to the two controls, is used to determine whether the two groups differ wherever it is appropriate to analyse individual sampling units. Where replicate means or medians are the unit for analysis, the Mann-Whitney rank sum test can be used. Further discussion of when each approach is appropriate is given in 5.2.2 and 5.2.2.4. 4.2.5 contains discussions of issues regarding multiple controls in an ecotoxicity study.

Figure 3 identifies a number of powerful methods for the analysis of quantal data. There are, of course, other statistical procedures that might be chosen. The following discussion identifies many of the procedures that might be used, gives details of some of the most appropriate, and attempts to provide some insight into the strengths and weaknesses of each method.

The methods used for determining NOEC values on quantal data can be categorized according to whether the tests involved are parametric or non-parametric and whether the methods are single-step or step-down. Table 3 lists methods that can be used to determine NOEC values. Some of these methods are applicable only under certain circumstances, and some methods are preferred over the others.

Except for the two Poisson tests, those tests listed in the column “Parametric” can be performed only when the study design allows the proportion of organisms responding in replicated experimental units to be

calculated (i.e. there are multiple organisms within each of multiple test vessels within each treatment group). Such a situation yields multiple responses, namely proportions, for each concentration, and these proportions can often be analysed as continuous. For very small samples, such a practice is inappropriate.

Typically, if responses increase or remain constant with increasing dosage, the trend-based methods perform better than pair-wise methods, and for most quantal data, a step-down approach based on the Cochran-Armitage test is the most appropriate of the listed techniques. The strengths and weaknesses of most listed methods are discussed in more detail below.

Table 3 — Methods used for determining NOEC values with quantal data

Method	Parametric	Non-Parametric
Single-step [a] (pair-wise)	Dunnett Poisson comparisons	Mann-Whitney [c] with Bonferroni-Holm correction Chi-squared with Bonferroni-Holm correction Steel's Many-to-One Fisher's Exact test with Bonferroni-Holm correction
Step-down [b] (trend-based)	Poisson trend Williams Bartholomew Welsch Brown-Forsythe Sequences of linear contrasts	Cochran-Armitage Jonckheere-Terpstra test Mantel-Haenszel
NOTE The tests listed in this table are well established as tests of the stated hypothesis in the statistics literature.		
[a] All listed single-step methods are based on pair-wise comparisons. [b] All step-down methods are based on trend tests. [c] The Mann-Whitney test is identical to the Wilcoxon rank-sum test.		

5.2.2 Parametric versus non-parametric tests

5.2.2.1 Basis

Parametric tests are based on assumptions that the responses being analysed follow some given theoretical distribution. Except for the Poisson methods, the tests listed in Table 3 as parametric all require that the data be approximately normally distributed (possibly after a transformation).The normality assumption can be met for quantal data only if the experimental design includes treatment groups that are divided into subgroups, the quantal responses are used to calculate proportions responding in each of the subgroups, and these proportions are the observations analysed. These proportions are usually subjected to a normalizing transformation (see 4.3.2, 4.3.3 and 4.3.4), and a weighted ANOVA is performed, perhaps with weights proportional to subgroup sizes (Cochran 1943). (It is noteworthy that some statistical packages, such as SAS version 6[7)], do not always perform multiple comparisons within a weighted ANOVA correctly.) This approach limits the possibilities of doing trend tests to those based on contrasts, including Welsch and Brown-Forsythe tests (Roth 1983, Brown and Forsythe 1974). Non-trend tests include versions of Dunnett's test for pair-wise comparisons allowing for unequal variances (Dunnett 1980, Tamhane 1979). These methods may not perform satisfactorily for quantal data, partly due to a loss of power in analysing subgroup proportions. An example is given in E.1.

The Cochran-Armitage test is listed as non-parametric, even though it makes explicit use of a presumed binomial distribution of incidence within treatment groups. Some reasons for this are given in E.1. Fisher's Exact test is likewise listed as non-parametric, even though it is based on the geometric distribution. The Jonckheere-Terpstra test applied to subgroup proportions is certainly non-parametric. An advantage of Jonckheere-Terpstra over the cited parametric tests is that the presence of many zeros poses no problem for

7) SAS version 6 is an example of a suitable product available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of this product.

the analysis and it provides a powerful step-down procedure in both large- and small-sample problems, provided the number of subgroups per concentration is not too small. An example in E.3 illustrates this concern.

5.2.2.2 Single-step procedures

Suitable single-step approaches for quantal data are Fisher's Exact test and the Mann-Whitney test to compare each treatment group to the control, independently of other treatment groups, with a Bonferroni-Holm adjustment. Details of these tests are given in E.1.

5.2.2.3 Step-down procedures

5.2.2.3.1 Choice of step-down procedure

Suitable step-down procedures for quantal data are based on the Cochran-Armitage and Poisson trend tests. First, a biological determination is made whether or not to expect a monotone dose response.

— If that judgement is to expect monotonicity, then the step-down procedure described below is followed, unless the data strongly indicate non-monotonicity.

— If the judgement is not to expect monotonicity, then Fisher's Exact test is used.

An analysis of quantal data is based on the relationships between the response (binary) variable and factors. In such cases, the Pearson Chi-squared (χ^2) test for independence can be used to find if any relationships exist.

5.2.2.3.2 Test for monotone dose response

If one believes on biological grounds that there is a monotone dose response, then the expected course of action is to use a trend test. However, statistical procedures should not be followed mindlessly. Rather, one should examine the data to determine whether it is consistent with the plan of action. There is a simple and natural way to check whether the dose response is monotone.

The $k-1$ df Pearson Chi-squared statistic decomposes into a test for linear trend in the dose response and a measure of lack of fit or lack of trend,

$$\chi^2_{(k-1)} = \chi^2_{(1)} + \chi^2_{(k-2)}$$

where

$\chi^2_{(1)}$ is the 1 df calculated Cochran-Armitage linear trend statistic; and

$\chi^2_{(k-2)}$ is the $k-2$ df Chi-squared test statistic for lack of fit.

The details of the computations are provided in E.1.

If the trend test is significant when all doses are included in the test, then proceed with a trend-based step-down procedure.

If the trend test with all doses included is not significant but the test for lack of fit is significant, then this indicates that there are differences among the dose groups but the dose response is not monotone. In this event, even if we expected a monotone dose response biologically, it would be unwise to ignore the contrary evidence and one should proceed with a pair-wise analysis.

The Cochran-Armitage trend test is available in several standard statistical packages including SAS and StatXact[8]. StatXact also provides exact power calculations for the Cochran-Armitage trend test with equally spaced or arbitrary doses.

5.2.2.3.3 Analysing the monotonic response for quantal data — Step-down procedure

A suitable approach to analysing the monotonic response for quantal data is as follows.

— Perform a Cochran-Armitage test for trend on responses from all treatment groups including the control.

— If the Cochran-Armitage test is significant at the 0,05 level, omit the high dose group, re-compute the Cochran-Armitage and Chi-squared tests with the remaining dose groups.

— Continue this procedure until the Cochran-Armitage test is first non-significant at the 0,05 level.

— The highest concentration remaining at this stage is the NOEC.

5.2.2.3.4 Possible modifications of the step-down procedure

There are two possible modifications to consider to the above.

First, as noted by Cochran (1943), Fisher's Exact test is more powerful for comparing two groups than the Cochran-Armitage test when the total number of subjects in the two groups is less than 20 and also when that total is less than 40 and the expected frequency of any cell is less than 5. This includes most laboratory ecotoxicology experiments. For this reason, if the step-down procedure described above reaches the last possible stage, where all doses above the lowest tested dose are significant, then we can substitute Fisher's Exact test for Cochran-Armitage for the final comparison on the grounds that it is a better procedure for this single comparison. Such substitution does not alter the power characteristics or theoretical justification of the Cochran-Armitage test for doses above the lowest dose, but it does improve the power of the last comparison.

Second, if the step-down procedure terminates at some higher dose because of a non-significant Cochran-Armitage test, but there is at this stage a significant test for lack of monotonicity, one should consider investigating the lower doses further. This can be done by using Fisher's Exact test to compare the remaining dose groups to the control, with a Bonferroni-Holm adjustment. The Bonferroni-Holm adjustment would take into account only the number of comparisons actually made using Fisher's Exact test. The inclusion of a method within the step-down procedure to handle non-monotonic results at lower doses is suggested for quantal data (but not for continuous data) for two reasons.

— First, there is a sound procedure built into the decomposition of the Chi-squared test for assessing monotonicity that is directly related to the Cochran-Armitage test.

— Secondly, experience suggests that quantal responses are more prone to unexpected changes in incidence rates at lower doses than continuous responses, so that a strict adherence to a pure step-down process may miss some adverse effects of concern.

5.2.2.4 Alternative procedures

5.2.2.4.1 Parametric and non-parametric procedures

The following parametric and non-parametric procedures are discussed because under some conditions, a parametric analysis of subgroup proportions may be the only viable procedure. This is especially true if there are also significant differences in the number of subjects within each subgroup, making analysis of means or medians problematic by other methods.

---

8) SAS and StatXact are examples of suitable products available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of these products.

5.2.2.4.2 Pair-wise ANOVA-based methods

Pair-wise ANOVA-based (weighted by subgroup size) methods performed on proportion affected have sometimes been used to determine NOEC values. While there can be problems with these proportion data meeting some of the assumptions of ANOVA (e.g. variance homogeneity), performing the analysis on proportion affected opens up the gamut of ANOVA-type methods, such as Dunnett's test and methods based on contrasts. Failure of data to satisfy the assumption of homogeneity of variances can often be corrected by the use of an arcsine-square-root or other normalizing and variance stabilizing transformation. However, this approach tends to have less power than step-down methods designed for quantal data that are described above, and is especially problematic for very small samples. These ANOVA-based methods may not be very powerful and are not available if there are not distinct subgroups of multiple subjects each within each concentration. Williams' test is a trend alternative that can be used, when data are normally distributed with homogeneous variance.

5.2.2.4.3 Jonckheere-Terpstra trend test

A non-parametric trend test that can be used to analyse proportion data is the Jonckheere-Terpstra trend test, which is intended for use when the underlying response on each subject is continuous and the measurement scale is at least ordinal. The most common application in a toxicological setting is for measures such as size, fecundity and time to an event. The details of this and other tests that are intended for use with continuous responses are given in 5.3. A disadvantage of the use of the Jonckheere-Terpstra trend test for analysing subgroup proportions where sample sizes are unequal is that it does not take sample size into account. It is not proper to treat a proportion based on 2 animals with the same weight as one based on 10, for example. For most toxicology experiments where survival is the endpoint, the sample sizes are equal, except for a rare lost subject, so this limitation is often of little importance. Where a sub-lethal effect on surviving subjects is the endpoint, then this is a more serious concern.

The methods described in Table 4 are sometimes used but tend to be less powerful than the ones designed for quantal data given in Table 3. They are appropriate only if responses of organisms tested are independent, and there is not significant heterogeneity of variances among groups (i.e. within-group variance does not vary significantly among groups). If there is a lack of independence or significant heterogeneity of variances, then modifications are needed. Some such modifications are discussed below. In the ANOVA context, a robust ANOVA (e.g. Welch's variance-weighted one-way ANOVA) that does not assume variance homogeneity can be used.

5.2.2.4.4 Poisson tests

Poisson tests can be used as alternatives in both non-trend and trend approaches (see Annex E.1). A robust Poisson approach (Weller and Ryan, 1998) using dummy variables for groups, or multiple Mann-Whitney tests using subgroup proportions as the responses could be used. In each case, an adjustment for number of comparisons should be made. For the robust Poisson model, this would be of the Bonferroni-Holm type. For the Mann-Whitney test, the Bonferroni-Holm adjustment could be used or these pair-wise comparisons could be "protected" by requiring a prior significant Kruskal-Wallis test (i.e. an overall rank-based test of whether any group differs from another). It should be noted that the Mann-Whitney approach does not take subgroup size into account, but this is usually not an issue for survival data.

5.2.2.5 Assumptions of methods for determining NOEC values

The assumptions that shall be met for the listed methods for determining NOEC values vary according to the methods. Assumptions common to all methods are given below, while others apply only to specific methods. The details on the latter are given in E.1.

Assumption: Responses are independent.

All methods listed in Table 4 are based on the assumption that responses are independent observations. Failure to meet this assumption can lead to highly biased results. If organisms in a test respond independently, they can be treated as binomially distributed in the analysis. (See 4.2.4 for further discussion.) It is not uncommon in toxicology experiments for treatment groups to be divided into subgroups.

For example, an aquatic experiment may have subjects exposed to the same nominal concentration but grouped in several different tanks or beakers. It sometimes happens that the survival rate within these subgroups varies more from subgroup to subgroup than would be expected if the chance of dying were the same in all subgroups. This added variability is known as extra-binomial (or extra-Poisson) variation, and is an indication that organisms in the subgroups are responding to different levels of an uncontrolled experimental factor (e.g. subgroups are exposed to differing light levels or are being held at differing temperatures) and are not responding independently. In this situation, correlations among subjects shall be taken into account.

For quantal responses, an appropriate way to handle this is to analyse the subgroup responses; that is, the subgroups are considered to be the experimental unit (replicate) for statistical analysis. Note that lack of independence can arise from at least two sources: differences in conditions among the tanks and interactions among organisms.

With mortality data, extra-binomial variation (heterogeneity) is not a common problem, but it is still advisable to do a formal or visual check. Two formal tests are suggested: a simple Chi-squared test and an improved test of Potthoff and Whittinghill (1966). Both tests are applied to the subgroups of each treatment group, in separate tests for each treatment group. While these authors do not suggest one, an adjustment for the number of such tests (e.g. Bonferroni) is advisable. It should be noted also that the Chi-squared test can become undependable when the number of expected mortalities in a Chi-squared cell is less than five. In this event, an exact permutation version of the Chi-squared test is advised and is available in commercially available software, such as StatXact and SAS.

If organisms are not divided into subgroups, lack of independence cannot be detected easily, and the burden for establishing independence falls to biological argument. If there is a high likelihood of aggression or competition between organisms during the test, responses may not be independent, and this possibility should be considered before assigning all organisms in a test level to a single test chamber.

It should be noted that even if subgroup information is entered separately, a simple application of the Cochran-Armitage test ignores the between-subgroup (i.e. within-group) variation and treats the data as though there were no subgrouping. This is inappropriate if heterogeneity among subgroups is significant. The same is true of simple Poisson modelling.

Thus, if significant heterogeneity is found, an alternative analysis is advised. One in particular deserves mention. This is a modification of the Cochran-Armitage test developed by Rao and Scott (1992) that is simple to use and is appropriate when there is extra-binomial variation. The beta-binomial model of Williams (1975) is another modification of the Cochran-Armitage tests that allows for extra-binomial variation. If the Jonckheere-Terpstra test is used, there is no adjustment (or any need to adjust) for extra-binomial variation, as that method makes direct use of the between-subgroup variation in observed proportions. However, as pointed out above, if there is considerable variation in subgroup sizes, this approach suffers by ignoring sample size.

Treatment of multiple controls

If both a solvent and water control are included in the experiment, a preliminary test can be done comparing just the two controls as a step in deciding how to interpret the experimental data. For quantal (e.g. mortality) data, Fisher's Exact test is appropriate. The decision of how to proceed after this comparison of controls is given in 4.2.5.

5.2.3 Additional information

E.1 contains details of the principle methods discussed in 5.2, including examples.

E.2 contains a discussion of the power characteristics of the step-down Cochran-Armitage test and Fisher's exact test.

5.3 and E.3 contain discussions of the methods for continuous responses that can be used to analyse subgroup proportions, as discussed above.

5.2.4 Statistical items to be included in the study report

The report describing quantal study results and the outcome of the NOEC determination should contain the following items:

— test endpoint assessed;

— number of test groups;

— number of subgroups within each group (if applicable);

— identification of the experimental unit;

— nominal and measured concentrations (if available) for each test group;

— number exposed in each treatment group (or subgroup if appropriate);

— number affected in each treatment group (or subgroup if appropriate);

— proportion affected in each treatment group (or subgroup if appropriate);

— confidence interval for the percent effect at the NOEC, provided that the basis for the calculation is consistent with the distribution of observed responses (see E.3);

— p value for test of homogeneity if performed;

— name of the statistical method used to determine the NOEC;

— the dose metric used;

— the NOEC;

— p value at the LOEC (if applicable);

— design power of the test to detect an effect of biological importance (and what that effect is) based on historical control background and variability;

— actual power achieved in the study;

— plot of response data versus concentration.

5.3 Hypothesis testing with continuous data (e.g. mass, length, growth rate) to determine NOEC

5.3.1 General

Figure 3 provides a scheme for determining NOEC values for continuous data, and identifies several statistical methods that are described in detail below. As reflected in this flow-chart, continuous monotone dose-response data are best analysed using a step-down test based on the Jonckheere trend test or Williams test (the former applicable regardless of the distribution of the data, the latter applicable only if data are normally distributed and variances of the treatment groups are homogeneous).

Non-monotonic dose-response data should be assessed using an appropriate pair-wise comparison procedure. Several such are described below. They can be categorized according to whether the data are normally distributed or homogeneous.

— Dunnett's test is appropriate if the data are normally distributed with homogeneous variance.

— For normally distributed but heterogeneous data, the Tamhane-Dunnett (T3) method (Hochberg and Tamhane 1987) can be used. Alternatively, such data can be analysed by the Dunn, Mann-Whitney, or unequal variance *t*-tests with a Bonferroni-Holm correction.

— Non-normal data can be analysed by using Dunn or Mann-Whitney tests with a Bonferroni-Holm correction.

— Normality can be formally assessed using the Shapiro-Wilk test (Shapiro and Wilk 1965).

— Homogeneity of variance is assessed by Levene's test (Box 1953).

— Dunn's test, if used, should be configured only to compare groups to control.

All of these procedures are discussed in detail below. Alternatives exist to these if the software used does not include these more desirable tests. For normality, the Anderson-Darling, Kolmogorov-Smirnov, Cramér-von Mises, Martinez-Iglewicz and D'Agostino Omnibus tests are available. For variance homogeneity, Cochran's Q, Bartlett's and the Maximum F tests can be used. The tests described in detail in this clause are recommended where available, based on desirable statistical properties.

There are, of course, a number of statistical procedures that are not listed in Figure 3 that might also be applied to continuous data. The following discussion identifies many of the procedures that might be used, and attempts to provide some insight into the strengths and weaknesses of each.

Table 4 lists methods that are sometimes used to determine NOEC values. Some of these methods are applicable only under certain circumstances, and some methods are preferred over the others. Parametric tests listed are performed only when the distribution of the data to be analysed is approximately normally distributed. Some parametric methods also require that the variances of the treatment groups be approximately equal.

Table 4 — Methods used for determining NOEC values with continuous data

Method	Parametric	Non-Parametric
Single-step [a] (pair-wise)	Dunnett Tamhane-Dunnett	Dunn Mann-Whitney with Bonferroni-Holm correction
Step-down [b] (trend-based)	Williams Bartholomew Welch trend Brown-Forsythe trend Sequences of linear contrasts	Jonckheere-Terpstra Shirley
[a] All listed single-step methods are based on pair-wise comparisons. [b] All step-down methods are based on trend tests.		

5.3.2 Parametric versus non-parametric tests

The parametric tests listed in Table 4, all require that the data be approximately normally distributed. Many also require that the variances of the treatment groups be equal (exceptions are the Tamhane-Dunnett, Welch and Brown-Forsythe tests). Parametric tests are desirable when these assumptions can be met. The failure of the data to meet assumptions can sometimes be corrected by transforming the data (see 4.3.4).

Some non-parametric tests are almost as powerful as their parametric counterparts when the assumptions of normality and homogeneity of variances are met. The non-parametric tests may be much more powerful if the assumptions are not met. Furthermore, a test based on trend is generally more powerful than a pair-wise test.

A decision to use a parametric or non-parametric test should be based on which best describes the physical, biological and statistical properties of a given experiment.

Piegorsch and Bailer (1997) warn that use of the Jonckheere-Terpstra test requires that shapes of distributions or the response variable be equivalent. In many cases, this translates to requiring that the response variable have a common variance. They conclude the applicability of the Jonckheere-Terpstra test is brought into question when there are large disparities in variances. While the Jonckheere-Terpstra test discussed in detail below is a distribution-free trend test, that fact alone does not mean that its results are not susceptible to heterogeneity of variance. While most people who have investigated the usual non-parametric methods find them less sensitive to these problems than the usual parametric procedures, they are not impervious to these problems. To address this question, a large power simulation study has been carried out (J.W. Green, manuscript in preparation) comparing the effects of variance heterogeneity on the Jonckheere, Dunnett, and Tamhane-Dunnett tests. These simulations have shown the Jonckheere test to be much less affected by heterogeneity than the alternatives indicated and to lose little of its good power properties.

Heterogeneity and non-normality are inherent in some endpoints, such as first or last day of hatch or swim up. There is observed zero within-group variance in the control and lower concentrations quite often and non-zero variance in higher concentrations. No transformation makes the data normal or homogeneous. It may be possible to apply some generalized linear model with a discrete distribution to such data, but that is not addressed in this clause.

5.3.3 Single-step (pair-wise) procedures

5.3.3.1 General

These tests are used when there is convincing evidence (statistical or biological) that the dose response is not monotone. This evidence can be through formal tests or through visual inspection of the data, as discussed in 5.3.4. Pair-wise procedures are also appropriate when there are differences among the treatments other than dose, such as different chemicals or formulations. These tests are described briefly here. Details of each test, including mathematical description, power, assumptions, advantages and disadvantages, relevant confidence intervals, and examples are discussed in E.3.

5.3.3.2 Dunnett's test

Dunnett's test is based on simple t-tests from ANOVA but uses a different critical value that controls the family-wise error (FWE) rate for the $k-1$ comparisons of interest at exactly α. Each treatment mean is compared to the control mean. This test is appropriate for responses that are normally distributed with homogeneous variances and is widely available.

5.3.3.3 Tamhane-Dunnett test

Also known as the T3 test, this is similar in intent to Dunnett's test but uses a different critical value and the test statistic for each comparison uses only the variance estimates from those groups. It is appropriate when the within-group variances are heterogeneous. It still requires within-group responses to be normally distributed and controls the FWE rate at exactly α.

5.3.3.4 Dunn's test

This non-parametric test is based on contrasts of mean ranks. In toxicity testing, it is used to compare the mean rank of each treatment group to the control. To control the FWE rate at α or less, the Bonferroni-Holm correction (or comparable alternative) should be applied. Dunn's test is appropriate when the populations have identical continuous distributions, except possibly for a location parameter (e.g. the group medians differ), and observations within samples are independent. It is used primarily for non-normally distributed responses.

5.3.3.5 Mann-Whitney test

This is also a non-parametric test and can be applied under the same circumstances as Dunn's test. The Mann-Whitney rank sum test compares the ranks of measurements in two independent random samples and has the aim of detecting if the distribution of values from one group is shifted with respect to the distribution of values from the other. It can be used to compare each treatment group to the control. When more than one comparison to the control is made, a Bonferroni-Holm adjustment is used.

5.3.4 Step-down trend procedures

For continuous data, two trend tests are described for use in step-down procedures, namely the Jonckheere-Terpstra test and Williams' test (described below), that are appropriate provided there is a monotone dose response. Where expert judgement is available, the assessment of monotonicity can be through visual inspection. For such an assessment, plots of treatment means, subgroup means, and raw responses versus concentration is helpful. An inspection of treatment means alone may miss the influence of outliers. However, a visual procedure cannot be automated, and some automation may be necessary in a high-volume toxicology facility. Although not discussed here in detail, the same methodology can be applied to the Welsch, Brown-Forsythe or Bartholomew trend tests.

A general step-down procedure is described in the next subclause. Where the term "trend test" is used, one may substitute either "Jonckheere-Terpstra test" or "Williams' test". Details of these, as well as advantages and disadvantages, examples, power properties, and related confidence intervals for each are given in E.3.

5.3.5 Determining the NOEC using a step-down procedure based on a trend test

5.3.5.1 General

This subclause describes a generalized step-down procedure for determining the NOEC for a continuous response from a dose-response study. It is appropriate whenever the treatment means are expected to follow a monotone dose response and there is no problem evident in the data that precludes monotonicity.

5.3.5.2 Preliminaries

The procedure described is suitable if the experiment being analysed is a dose-response study with at least two dose groups (Figure 8). For clarity, the term "dose group" includes the zero-dose control. Before entering the step-down procedure, two preliminary actions shall be taken.

First, the data are assessed for monotonicity (as discussed in 5.1.5). A step-down procedure based on trend tests is used if a monotonic response is evident. Pair-wise comparisons (e.g. Dunnett's, Tamhane-Dunnett, Dunn's test or Mann-Whitney with a Bonferroni-Holm correction, as appropriate) instead of a trend-based test should be used where there is strong evidence of departure from monotonicity.

Next, examine the number of responses and number of ties (as discussed in 5.3.6.1). Small samples and data sets with massive ties should be analysed using exact statistical methods if possible. Finally, if a parametric procedure (e.g. Dunnett's or Williams' test) is to be used, then an assessment of normality and variance homogeneity should be made. These are described in 5.3.6.2 and 5.3.6.3, respectively.

5.3.5.3 Step-down procedure

5.3.5.3.1 Preferred approach

The preferred approach to analysing monotonic response patterns is as follows.

— Perform a test for trend (Williams or Jonckheere) on responses from all dose groups including the control.

— If the trend test is significant at the 0,05 level, omit the high dose group, and re-compute the trend statistic with the remaining dose groups.

— Continue this procedure until the trend test is first non-significant at the 0,05 level, then stop. The NOEC is the highest dose remaining at this stage.

— If this test is significant when only the lowest dose and control remain, then an NOEC cannot be established from the data.

5.3.5.3.2 Williams' test

Williams' test is a parametric procedure that is applied in the same way the Jonckheere-Terpstra test is applied. This procedure, described in detail in E.3, assumes data within concentrations are normally distributed and homogeneous.

In addition to the requirement of monotonicity rather than linearity in the dose response, an appealing feature of this procedure is that maximum likelihood methods are used to estimate the means (as well as the variance) based on the assumed monotone dose response of the population means. The resulting estimates are monotone.

An advantage of this method is that it can also be adapted to handle both between- and within-subgroup variances. This is important when there is greater variability between subgroups than chance alone would indicate. Williams' test shall be supplemented by a non-parametric procedure to cover non-normal or heterogeneous cases. Either Shirley's (1979) non-parametric version of Williams' test or the Jonckheere-Terpstra test can be used, but if these alternative tests are used, one loses the ability to incorporate multiple sources of variances. Limited power comparisons suggest similar power characteristics for Williams' and the Jonckheere-Terpstra tests.

5.3.5.3.3 Jonckheere-Terpstra test

The Jonckheere-Terpstra trend test is intended for use when the underlying response of each experimental unit is continuous and the measurement scale is at least ordinal.

The Jonckheere-Terpstra test statistic is based on joint rankings (also known as Mann-Whitney counts) of observations from the experimental treatment groups. These Mann-Whitney counts are a numerical expression of the differences between the distributions of observations in the groups in terms of ranks. The Mann-Whitney counts are used to calculate a test statistic that is used in conjunction with standard statistical tables to determine the significance of a trend. E.3 gives details of computations. The Jonckheere-Terpstra test reduces to the Mann-Whitney test when only one group is being compared to the control.

The Jonckheere-Terpstra test has many appealing properties. Among them is the requirement of monotonicity rather than linearity in the dose response. Another advantage is that an exact permutation version of this test is available to meet special needs (as discussed below) in standard statistical analysis packages, including SAS and StatXact. If subgroup means or medians are to be analysed, the Jonckheere-Terpstra test has the disadvantage of failing to take the number of individuals in each subgroup into account.

Extensive power simulations of the step-down application of the Jonckheere-Terpstra test compared to Dunnett's test have demonstrated in almost every case considered where there is a monotone dose response, that the Jonckheere-Terpstra test is more powerful than Dunnett's test (Green J.W., in preparation for publication). The only situation investigated in which Dunnett's test is *sometimes* slightly more powerful than the Jonckheere-Terpstra is when the dose response is everywhere flat except for a single shift. These simulations followed the step-down process to the NOEC determination by the rules given above and covered a range of dose-response shapes, thresholds, number of groups, within-group distributions, and sample sizes. The development of the Jonckheere-Terpstra test is available in many places, including Lehmann (1975).

5.3.6 Assumptions for methods for determining NOEC values

5.3.6.1 Small samples — Massive ties

Many standard statistical tests are based on large-sample or asymptotic theory. If a design calls for fewer than 5 experimental units per concentration, such large-sample statistical methods may not be appropriate. In

addition, if the measurement is sufficiently crude, then a large proportion of the measured responses have the same value, or are very restricted in the range of values, so that tests based on a presumed continuous distribution may not be accurate. In these situations, an exact permutation-based methodology may be appropriate.

While universally appropriate criteria are difficult to formulate, a simple rule that should flag most cases of concern is to use exact methods when any of the following conditions exists:

a) at least 30 % of the responses have the same value;

b) at least 50 % of the responses have one of two values;

c) at least 65 % of the responses have one of three values.

StatXact and SAS are readily available software packages that provide exact versions of many useful tests, such as the Jonckheere-Terpstra and Mann-Whitney tests.

5.3.6.2 Normality

When parametric tests are being considered for use, then a Shapiro-Wilk test (Shapiro and Wilk 1965) of normality should be performed. If the data are not normally distributed, then either a normalizing transformation (4.3.4) should be sought or a non-parametric analysis should be done. Assessment of non-normality can be done at the 0,05 significance level, though a 0,01 level might be justified on the grounds that ANOVA is robust against mild non-normality. The data to be checked for normality are the residuals after differences in group means are removed; for example, from an ANOVA with concentration, and, where necessary, subgroup, as class (i.e. non-numeric) variables.

5.3.6.3 Variance homogeneity

If parametric tests are being considered for use and the data are normally distributed, then a check of variance homogeneity should be performed. Levene's test (Box, 1953) is reasonably robust against marginal violations of normality. If there are multiple subgroups within concentrations, the variances used in Levene's test are based on the subgroup means. If there are no subgroups, the variances based on individual measurements within each treatment group would be used. It should be noted that ANOVA is robust to moderate violations of assumptions, especially if the experimental design is balanced (equal n in the treatment groups), and that some tests for homogeneity are less robust than the ANOVA itself. Small departures from homogeneity (even though they may be statistically significant by some tests) can be tolerated without adversely affecting the power characteristics of ANOVA-based tests. For example, it is well known that Bartlett's test is very sensitive to non-normality. It is customary to use a much smaller significance level, (e.g. 0,001) if this test is used. Levene's test, on the other hand, is designed to test for the very departures from homogeneity that cause problems with ANOVA, so that a higher level significance (0,01 or 0,05) in conjunction with this test can be justified. Where software is available to carry out Levene's test, it is recommended over Bartlett's.

For pair-wise (single-step) procedures, if the data are normally distributed but heterogeneous, then a robust version of Dunnett's test (called Tamhane-Dunnett in this Technical Specification) is available. Such a procedure is discussed in Hochberg and Tamhane (1987). Alternatives include the robust pair-wise tests of Welch and Brown-Forsythe. If the data are normally distributed and homogeneous, then Dunnett's test is used. Specific assumptions and characteristics of many of the tests referenced in this subclause are given in E.3.

Of course, expert judgement should be used in assessing whether a significant formal test for normality or variance homogeneity reveals a problem that calls for alternative procedures to be used.

5.3.7 Operational considerations for statistical analyses

5.3.7.1 Treatment of experimental units

A decision that shall often be made is whether the individual animals or plants can be used as the experimental unit for analysis, or whether subgroups should be the experimental unit. The consequences of this choice should be carefully considered. If there are subgroups in each concentration, such as multiple tanks or beakers or pots, each with multiple specimens, then the possibility exists of within- and among-subgroup variation, neither of which should be ignored. If subjects within subgroups are correlated, that does not mean that individual subject responses should not be analysed. It does mean that these correlations should be explicitly modelled or else analysis should be based on subgroup means. Methods for modelling replicated dose groups (e.g. nested ANOVA) are available. For example, Hocking (1985), Searle (1987, especially 13.5), Milliken and Johnson (1984, especially Chapter 23), John (1971), Littell (2002) and many additional references contain treatments of this.

Technical note: If both within-subgroup and between-subgroup variations exist and neither is negligible, then the step-down trend test should either be the Jonckheere-Terpstra test with mean or median subgroup response as the observation, or else an alternative trend test such as Williams' or Brown-Forsythe with the variance used being the correct combination of the within- and among-subgroups variances as described in the discussion on the Tamhane-Dunnett test in E.3.

Given the possibility of varying subgroup sample sizes at the time of measurement, it may not be appropriate to treat all subgroup means or medians equally. For parametric comparisons, this requires only the use of the correct combination of variance components, again as described as E.3. For non-parametric methods, including Jonckheere's test, there are no readily available methods for combining the two sources of variability. The choices are between ignoring the differences in sample sizes and ignoring the subgroupings. If the differences in sample sizes are relatively small, they can be ignored. If the differences among subgroups are relatively small, they can be ignored. If both differences are relatively large, then there is no universal best method. A choice can be made based on what has been observed historically in a given laboratory or for a given type of response and built into the decision tree.

5.3.7.2 Identification and meaning of outliers

The data should be checked for outliers that might have undue influence on the outcome of statistical analyses. There are numerous outlier rules that can be used. Generally, an outlier rule such as Tukey's (Tukey, 1977) that is not itself sensitive to the effects of outliers is preferable to methods based on standard deviations, which are quite sensitive to the effects of outliers. Tukey's outlier rule can be used as a formal test with outliers being assessed from residuals (results of subtracting treatment means from individual values) to avoid confounding outliers and treatment effects.

Any response more than 1,5 times the interquartile range above the third quartile (75th percentile) or below the first quartile (25th percentile) is considered an outlier by Tukey's rule. Such outliers should be reported with the results of the analysis. The entire analysis of a given endpoint can be repeated with outliers omitted to determine whether the outliers affected the conclusion. While it is true that non-parametric analyses are less sensitive to outliers than parametric analyses, omission of outliers can still change conclusions, especially when sample sizes are small or outliers are numerous.

Conclusions that can be attributed to the effect of outliers should be carefully assessed. If the conclusions are different in the two analyses, a final analysis using non-parametric methods may be appropriate, as they are less influenced than parametric methods by distributional or outlier issues.

It is not appropriate to omit outliers in the final analysis unless this can be justified on biological grounds. The mere observation that a particular value is an outlier on statistical grounds does not mean it is an erroneous data point.

5.3.7.3 Multiple controls

To avoid complex decision rules for comparing a water and solvent control, it is recommended that a non-parametric Mann-Whitney (or, equivalently, Wilcoxon) comparison of the two controls be performed, using

only the control data. This comparison can be either a standard or an exact test, according to whether the preliminary test for exact methods is negative or positive. If a procedure for comparing controls using parametric tests were to be employed, then another layer of complexity can result, where one has to assess normality and variance homogeneity twice (once for controls and again later, for all groups) and one shall also consider the possibility of using transformations in both assessments.

5.3.7.4 General

Outliers, normality, variance homogeneity and checks of monotonicity should be done only on the full data set, not repeated at each stage of the step-down trend test, if used. Diagnostic tools for determining influential observations can also be very helpful in evaluating the sensitivity of an analysis to the effects of a few unusual observations.

5.4 Statistical items to be included in the study report

The report describing continuous study results and the outcome of the NOEC determination should contain the following items:

— description of the statistical methods used;

— test endpoint assessed;

— number of test groups;

— number of subgroups within each group and how handled (if applicable);

— identification of the experimental unit;

— nominal and measured concentrations (if available) for each test group;

— the dose metric used;

— number exposed in each treatment group (or subgroup if appropriate);

— group means (and median, if a non-parametric test was used) and standard deviations;

— confidence interval for the percent effect at the NOEC, provided that the basis for the calculation is consistent with the distribution of observed responses (see E.3);

— the NOEC;

— p value at the LOEC (if applicable);

— results of power analysis;

— plot of response versus concentration.

6 Dose-response modelling

6.1 Introduction

The main regulatory use of dose-response modelling in toxicity studies is to estimate an EC_x, the exposure concentration that causes an x % effect in the biological response variable of interest, and its associated confidence bounds. The value of x, the percent effect, may be specified in advance, based on biological (or regulatory) considerations. Guidelines may specify for which value(s) of x the EC_x is required. This clause discusses how an EC_x may be estimated, as well as how it may be judged that the available data are sufficient to do so.

Dose-response (or concentration-response) modelling aims at describing the dose-response data as a whole, by means of a dose-response model. In general terms, it is assumed that the response, y, can be described as a function of concentration (or dose), x:

$$y = f(x)$$

where f can be any function that is potentially suitable for describing a particular dataset.

Since y is considered as a function of x, the response variable y is also called the dependent variable, and the concentration x, the independent variable. As an example, consider the linear function:

$$y = a + b\,x$$

where the response changes linearly with the concentration. Here, a and b are called the model parameters.

By changing parameter a one may shift the line upwards or downwards, while by changing the parameter b one may rotate the line. Fitting a line to a dataset is the process of finding those values of a and b that result in "the best fit", i.e. making the distances of the data points to the line as small as possible. Similarly, for any other dose-response model, or function f, the best fit may be achieved by adjusting the model parameters.

This example illustrates that the data determine the values of the parameters a and b, and thereby the location and angle of the line. However, whatever the data, the result of the fitting process is, for this model, always a straight line, so the flexibility of the dose-response model in following the dose-response data is limited. In general, the flexibility of a dose-response model tends to be larger when it includes more parameters. For example, the model:

$$y = a + b\,x + c\,x^2 + d\,x^3$$

has four parameters (a, b, c and d), which can all be varied in the fitting process. Therefore, this model is more flexible compared to the linear model, and can take on various shapes other than a straight line. One might conclude here: "the more parameters, the better", but that is not the case. It only makes sense to include more parameters in a model when the data contain the information to estimate them [also referred to as the parsimony principle (3.25)], or when including the parameter in the model leads to a significantly better fit.

The fit of the model to the data may be defined in various ways. One measure for the fit is the sum of squares of the residuals, where the residuals are simply the distances (differences) between the data and the model value at the pertinent concentration. The best fit is then found by minimizing the sum of squared residuals, or briefly the Sum of Squares (SS). Another measure for the fit is the likelihood, which is based on a particular distribution that is assumed for the data (e.g. a normal or log-normal distribution for continuous data, a binomial distribution for quantal data, or a Poisson distribution for count data). In that case, the best fit is found by maximizing the likelihood (or the log-likelihood, which amounts to the same thing). See 4.3.6. for a general discussion of model fitting.

In this clause, a dose-response model is generally written as $y = f(x)$, where x may denote either the concentration or dose. Indeed, a concentration-response and a dose-response model are not different from a statistical point of view. The response, y, may refer to data of various types. The type of the response data, either quantal or continuous (see Clause 3), does make an important difference, not only for the statistical

analysis, but also for the interpretation of the results. In this clause, dose-response modelling is separately discussed for quantal (6.2) and for continuous (6.3) data, since the statistical analysis is completely different. The flow-chart given in Figure 7 summarizes the main lines of a dose-response modelling approach.

Of course, the response in biological test systems not only depends on the concentration (dose) but also on the exposure duration. Yet, most ecotoxicity tests only vary the concentration (dose), at a single exposure duration. Therefore, the larger part of this clause addresses how to model the concentration-response relationship, ignoring the exposure duration. Obviously, any results from the statistical analysis then only hold for that particular exposure duration.

For data sets where the exposure duration varied as well, one may apply models where the response is a function of both concentration and exposure duration. The inclusion of exposure duration is discussed in 6.6, for both quantal and continuous data. In the other subclauses of this clause, time is considered as fixed. In Clause 7, the role of time and exposure duration in describing the response is further discussed from the perspective of biology-based modelling.

6.2 Modelling quantal dose-response data (for a single exposure duration)

6.2.1 General

A quantal response, y, is defined as $y = k/n$, where k is the number of responding organisms (or experimental units) out of a total of n.

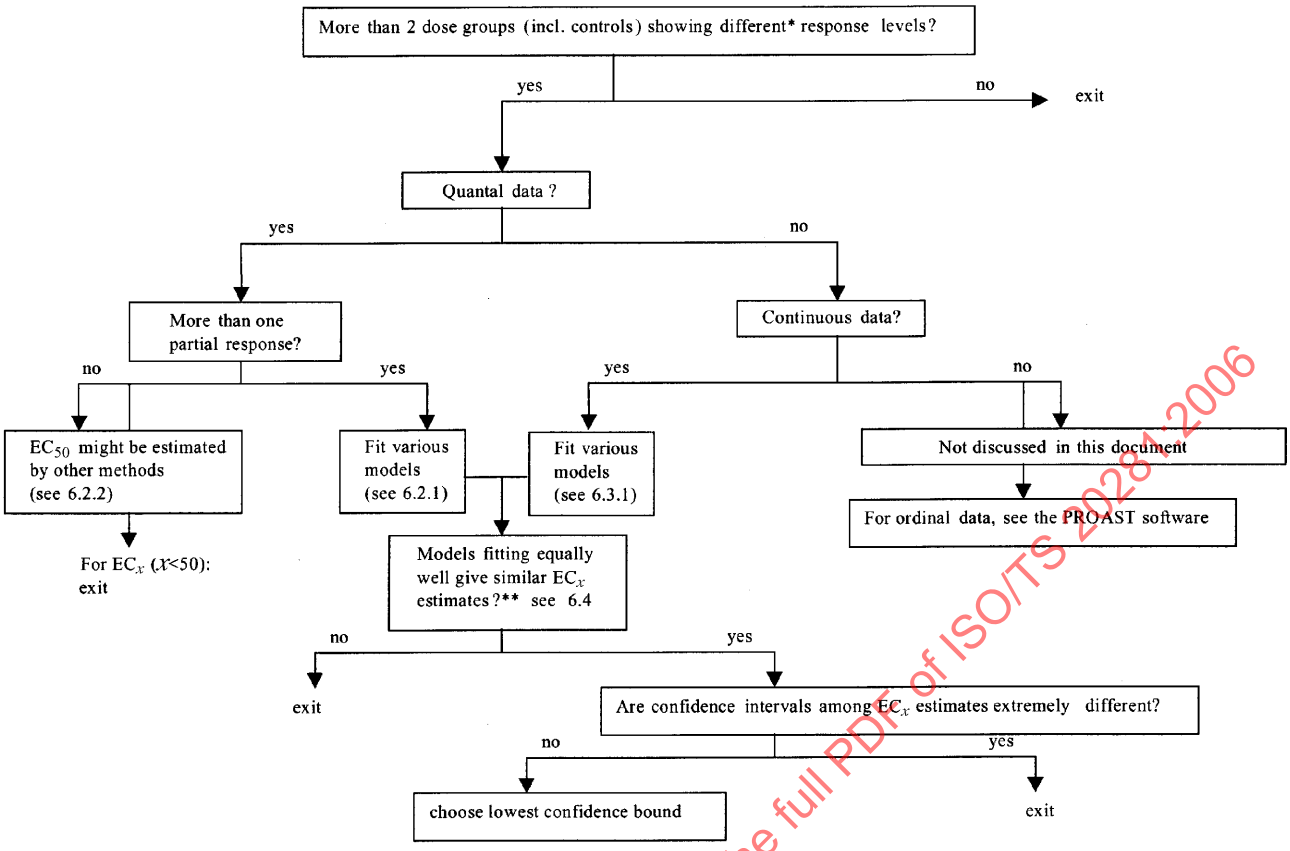
A quantal response may also be expressed as a percentage, but the total number of observed units, n, cannot be omitted. For example, 2 responses out of 4 is not the same information as 50 out of 100. See also Clause 4.

The purpose of dose-response modelling of quantal data is to estimate an EC_x (LC_x), where x is a given percentage, usually equal to or lower than 50 %. When the dose-response data relate to a particular (single) exposure duration, the estimated parameters (EC_{50} or EC_x) obviously only hold for that particular exposure duration (or to, e.g. a single acute oral dose).

In this clause, the terms $ED_{50}/EC_{50}/LD_{50}/LC_{50}$ are used interchangeably, as well as $ED_x/EC_x/LD_x/LC_x$, where x denotes a particular response level (usually smaller than 50). Note that x in model expressions denotes the concentration (or dose). In human risk assessment, the term Benchmark dose (BMD)[9] is used, and is, for the case of quantal data, equivalent to the EC_x (but not so for continuous responses, see 6.3).

The terms dose and concentration are used interchangeably, as well as dose-response relationship and concentration-response relationship.

9) Orginally the BMD was introduced by Crump (1984) as the lower confidence bound of the point estimate. More recently this is often indicated as BMDL.



Key

Exit: EC_x cannot be assessed from the data at hand; repeat the experiment with more (adequate) doses, or go to Clause 5.

* I.e. apparently different, given the noise in the data.

** In addition, it should be assessed by visual inspection that the fitted model is sufficiently supported by the data.

NOTE Doses = concentrations.

Figure 6 — Flow-chart for dose-response modelling

6.2.2 Choice of model

6.2.2.1 General

A (statistical) dose-response model serves to express the observed response as a function of dose, to provide for a tool to estimate the parameters of interest (in particular the EC_x) and assess confidence intervals for those estimates. A statistical regression model itself does not have any meaning, and the choice of the model (expression) is largely arbitrary. It is the data, not the model, that determines the dose response, and thereby the EC_x. Of course, an improper choice of the model can lead to an inappropriate estimate of the EC_x, but the choice of the model is in most ecotoxicity studies governed by the data.

Numerous dose-response models are theoretically possible, but in practice only a limited number are applied, mostly determined by historical habits in the field of application. Only the more frequently applied models are discussed here. See 6.4 for a discussion on model selection.

For quantal data, an obvious property for a dose-response function is that it ranges between 0 and 1 (0 % and 100 %). Further, one would normally expect the response to be monotone, i.e. it only increases (or decreases). Cumulative distribution functions (e.g. normal, logistic, Weibull) obey that property, and are therefore candidates for dose-response modelling of quantal data.

The use of cumulative distribution functions for quantal dose-response modelling can also be considered from the idea of tolerance distributions. By assuming that each individual in the population observed has its own tolerance for the chemical, a tolerance distribution expresses the variability between the individuals. Plotting the tolerance distribution cumulatively results in the quantal dose-response relationship, where the fraction of responding individuals (at a given concentration) is viewed as all individuals having a tolerance lower than that concentration. For example, a predicted response of 25 % at concentration 10 ppm is interpreted as 25 % of the individuals having a tolerance lower than 10 ppm. Given this interpretation, the slope of a quantal dose-response relationship is a reflection of the variability between the individuals, with steeper slopes meaning smaller variability in tolerances.

In light of the preceding, the choice of a quantal concentration-response model may be based on an assumed tolerance distribution. For several reasons, one may expect a tolerance distribution to be approximately log-normal, or, equivalently, to be approximately normal for the log-concentrations. Indeed, a long history of experience has confirmed this, and it has become standard that models that are based on symmetrical tolerance distributions (e.g. the probit and logit models, see below) are fitted against the logarithms of the concentrations.

In general, a dose-response model for quantal data is a function of the concentration or dose x:

$$y = f(x)$$

where y is the quantal response.

It is important to keep in mind that in the model, y represents the true response, which may be thought of as the fraction of responding individuals in the *infinite* population, or as the probability of response for any individual. The function $f(x)$ is chosen such that it equals zero at concentration zero (and unity for infinite concentration). However, theoretically, the probability of response in the unexposed population might be very small, but it cannot be (strictly) zero. Therefore, it is theoretically more appropriate to extend the model and include a background incidence parameter by putting

$$y = f(x) = a + (1 - a)\, g(x)$$

where a denotes the true background probability of response, and $g(x)$ is a function increasing from 0 to 1 for x increasing from zero to infinity. In this formulation, the response at infinite concentrations remains unity (since $g(x) = 1$ for infinite x).

Some of the more commonly used models are discussed below. For a more extensive list of models, see Scholze *et al*. (2001).

6.2.2.2 Probit model

The probit model is the cumulative normal distribution function. In practice, it is usually applied to the log-concentrations, implying that a log-normal tolerance distribution is assumed. The probit model (without the background mortality parameter a) can be expressed as

$$z(y) = b\,\left[\log(x) - \log(\mathrm{ED}_{50})\right] = b\,\log\left(\frac{x}{\mathrm{ED}_{50}}\right) \qquad (1)$$

where z is the standard normal deviate associated with probability y.

At first sight, the use of log-concentration in this model appears to present a problem for dose zero. Note, however, that for

$x = 0$, $z = -\infty$, and the associated probability, y, is zero.

In other words, Equation (1) assumes that the probability of ever observing a response in the control group is strictly zero. Therefore, when Equation (1) is fitted to the data, the control observations can just as well be

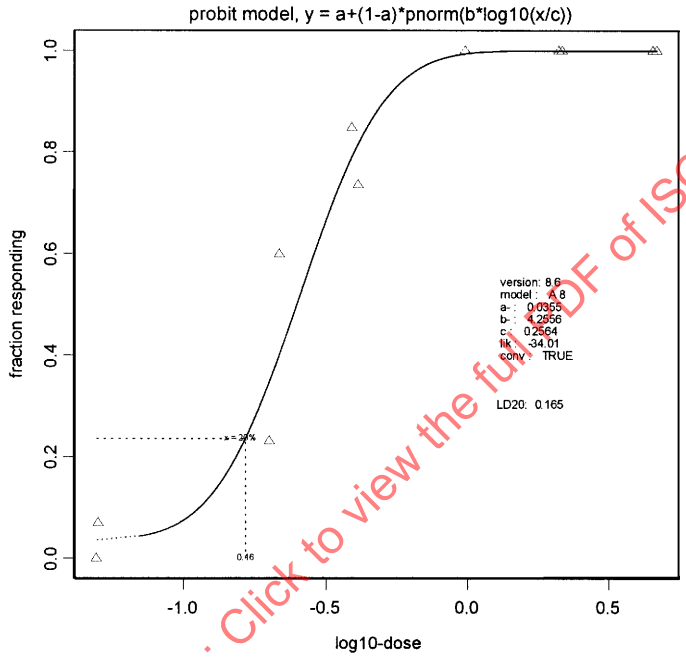
deleted[10]. They only provide information to the model when a background parameter is included in the model [see Equation (5)].

The standard normal deviate cannot be calculated from an explicit expression, as opposed to the logit model (see below). Common statistical software packages use standard algorithms; therefore, this should not concern the user.

The probit model has two parameters: the ED_{50} and the slope (b).

The ED_{50} is the median of the (log-normal) tolerance distribution, and the slope is the inverse of the standard deviation of that distribution.

Figure 7 shows an application of the probit model to mortality data.



Key

a background mortality

b slope

c LD_{50}

NOTE Dashed lines indicate the LD_{20}.
"pnorm" is the cumulative standard normal distribution function.
On log-scale, the zero concentration is minus infinity.

Figure 7 — Probit model fitted to observed mortality frequencies (triangles) as a function of log-dose

10) Adequate software simply sets the log-likelihood score for observations in the control group at zero (whatever the observations). When a background response parameter is included in the model [see Equation (5)], the log-likelihood score associated with the observations in the control group only depends on the value of the background parameter.

6.2.2.3 Logit model

The logit model is the cumulative logistic distribution function. The logistic distribution has wider tails than the normal distribution, but is similar otherwise. Just as with the probit model, the logit model is usually applied to the log-concentrations.

The logit model (without the background mortality parameter, a) can be expressed as

$$y = \frac{1}{1 + \exp[b \log(\mathrm{ED}_{50}/x)]} \tag{2}$$

where y is the probability of response.

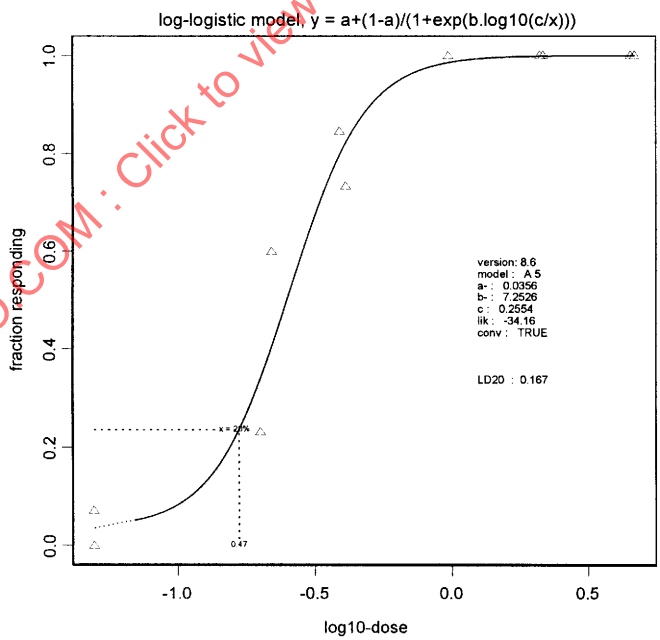
Just as in the probit model, the logarithm of dose does not present any problem for dose zero. It can be seen immediately that, in the limit, y equals zero for x approaching zero. In fitting the model, the control observations can be simply deleted, as they do not provide any information, unless a background parameter is included [see Equation (5)].

The logit model has two parameters: the ED_{50} and the slope, b.

The ED_{50} is the median of the (log-logistic) tolerance distribution, and the slope is related to the standard deviation, s, by

$$s_{\mathrm{tol,dist}} = \frac{\pi}{b\sqrt{3}}$$

Figure 8 illustrates the logit model applied to the same mortality data as Figure 7.



Key

a background mortality

b slope

c LD_{50}

NOTE Dashed lines indicate the LD_{20}.

Figure 8 — Logit model fitted to mortality dose-response data (triangles)

6.2.2.4 Weibull model

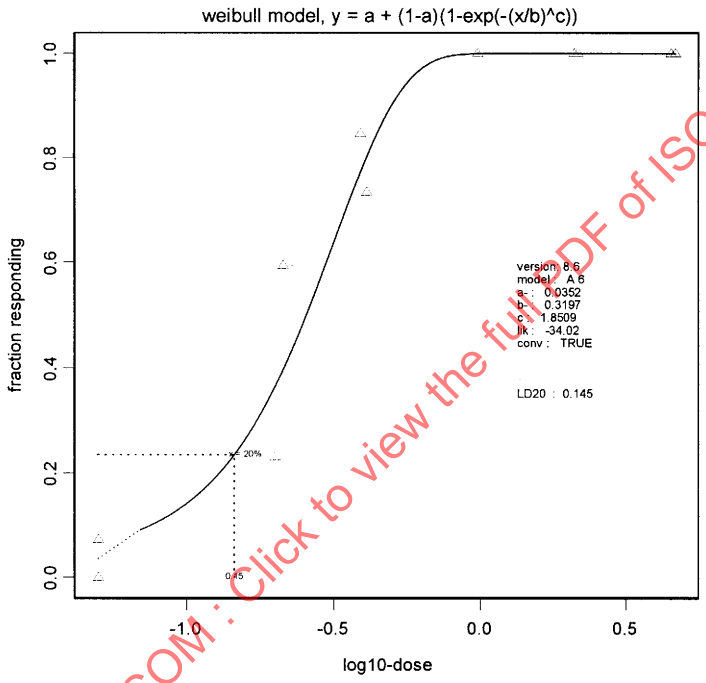
The Weibull distribution is not necessarily symmetrical, and is usually applied to the concentrations themselves (not their logs). The Weibull model (without the background mortality parameter, a) may be expressed as

$$y = 1 - \exp\ [\ -\ (x/b)^c\ ] \quad (3)$$

It has two parameters, a location parameter, b, and a parameter c (high values of c give steep slope). The ED_{50} is related to b and c by

$$ED_{50} = b\ \ln(2)^{1/c}$$

Figure 9 illustrates the Weibull model applied to the same mortality data as Figures 7 and 8.



Key

a background mortality

b "location" parameter

c "slope" parameter

NOTE Dashed lines indicate the LD_{20}.

Figure 9 — Weibull model fitted to mortality dose-response data (triangles)

The LD_{50} equals $b\ \ln(2)^{1/c} = 0{,}145$.

NOTE The data and the model in Figures 7 to 9 are plotted against log-dose with the purpose of improving the readability of the plots. However, the Weibull model was fitted as a function of dose, while the probit and logit models were fitted as a function of log-dose.

6.2.2.5 Multi-stage models

The multi-stage model (see e.g. Crump *et al*., 1976) is often used for describing tumour dose-response data. It is usually applied in a simplified version (the linearized multi-stage model, briefly LMS):

$$y = 1 - \exp\{-a - bx - cx^2 - dx^3 - \ldots.\} \quad (4)$$

where the number of parameters is also called the number of stages. It includes the one-stage model:

$$y = 1 - \exp\{-a - bx\},$$

also referred to as the one-hit model.

Note that in the multi-stage model background mortality is included, and equals 1 – exp(-a).

The multi-stage model can be regarded as a family of nested models. For example, by setting the parameter d in the three-stage model equal to zero, one obtains the two-stage model. Thus, one can let the number of stages depend on the data (see below for a further discussion of nested models).

6.2.2.6 Definitions of EC_{50} and EC_x

The EC_x is defined as the concentration associated with x % response, with the EC_{50} as a special case[11]. The situation of nonzero background response complicates the definition of the EC_{50} and of the EC_x, since the background response may be taken into account in various ways.

The EC_{50} is defined as the concentration associated with 50 % response. However, a 50 % response (i.e. incidence) can relate to the whole population, irrespective of the background response, or only to that part of the population that did not respond at concentration zero. Consider the general quantal dose-response model where the background response (incidence), a, is included as a model parameter:

$$y = f(x) = a + (1 - a)\, g(x) \quad (5)$$

Here $g(x)$ may be any cumulative tolerance distribution, ranging from zero to one. It reflects the dose-response relationship for the fraction of the population that did not show a response at concentration zero. The background-corrected EC_{50} then simply is the EC_{50} as given by $g(x)$. For example, when $g(x)$ denotes the log-logistic model, then parameter c is the background-corrected EC_{50}. This definition of the EC_{50} (LD_{50}) is used in Figures 7 to 9.

For response levels x % smaller than 50 %, the EC_x may be defined in various ways, e.g.

x %/100 % = $f(EC_x) - a$ (additional risk)

x %/100 % = $[f(EC_x) - a] / (1 - a) = g(EC_x)$ (extra risk)

EXAMPLE 1 Additional risk definition

For example, when the background response, a, amounts to 3 %, then the EC_{10} according to the additional risk definition corresponds to a response in the population of 13 %, since 13 % – 3 % = 10 %.

EXAMPLE 2 Extra risk definition

In the extra risk definition, the EC_{10} would correspond to a response of 12,7 %, since (12,7 % – 3 %) / 97 % = 10 %.

11) In human risk assessment, the term Benchmark dose (BMD) is used, defined as the dose associated with a certain Benchmark response (= x %). Originally the Benchmark dose was defined as the lower confidence limit of the point estimate (Crump, 1984), also indicated as BMDL. See also Table 5 in 6.3.

Note that the background-corrected EC_x, according to the extra risk concept, is equal to the (uncorrected) EC_x of $g(x)$ in Equation (5). Therefore, extra risk appears favourable, but the numerical difference for the EC_x based on additional or extra risk is usually small. The illustrative examples in Figures 7 to 9 used the additional risk concept.

In ecotoxicity testing, the additional risk is common for the EC_x when $x < 50$ %. However, in the case of the EC_{50}, the background response is usually taken into account according to the extra risk concept (as in Figures 7 to 9).

It may be noted that in other disciplines, still other risk concepts are used. For instance, in epidemiology, more common measures are relative risk (response of exposed subjects divided by response in non-exposed subjects) and derived concepts, such as attributable proportion, and odds ratio.

6.2.3 Model fitting and estimation of parameters

6.2.3.1 Software and assumptions

Fitting a model to dose-response data may be done by using any suitable software, e.g. SAS (www.sas.com), SPSS (www.spss.com), splus (www.insightful.com)[12], and PROAST[13] (Slob, 2003).

The user does not need to be aware of the computational details, but some understanding of the basic principles in non-linear regression is required to be able to interpret the results properly. These principles are discussed in 6.7.

Furthermore, the user should be aware of the assumptions underlying the fit algorithm. For quantal data, it is usually assumed that the data follow a binomial distribution, and the common fit algorithm is based on maximizing the binomial likelihood (see 6.2.4 for a discussion of the assumptions). The parameter values produced by this algorithm are the values associated with the maximum likelihood, and are also called the Maximum Likelihood Estimates (MLEs).

Maximum likelihood can only be applied for data including at least two concentrations with partial responses, otherwise the MLE of the slope tends to infinity. When the data only include 0 % and 100 % responses, or only a single concentration with partial response, the slope of the dose response can therefore not be estimated. But there are several methods available for estimating the EC_{50} in those situations. These methods include procedures for assessing the precision of the estimated EC_{50} (Hoekstra, 1993).

6.2.3.2 Response in controls

Instead of estimating the background response (incidence) as a parameter in the dose-response model, Abbott's correction is often used in situations where dose-response data show nonzero observed response in the controls. In this correction, each observed response, p_i, is replaced by

$$(p_i - p_0) / (1 - p_0),$$

where p_0 denotes the observed background response.

However, this is inappropriate, since the observed background response, p_0, contains error, which is not taken into account in this way. Instead the background response should be treated as an estimate containing error, just like the observed responses in the other dose groups. By incorporating the background response as a parameter in the model, it is estimated from the data, and estimation errors are accounted for, e.g. in calculating confidence intervals.

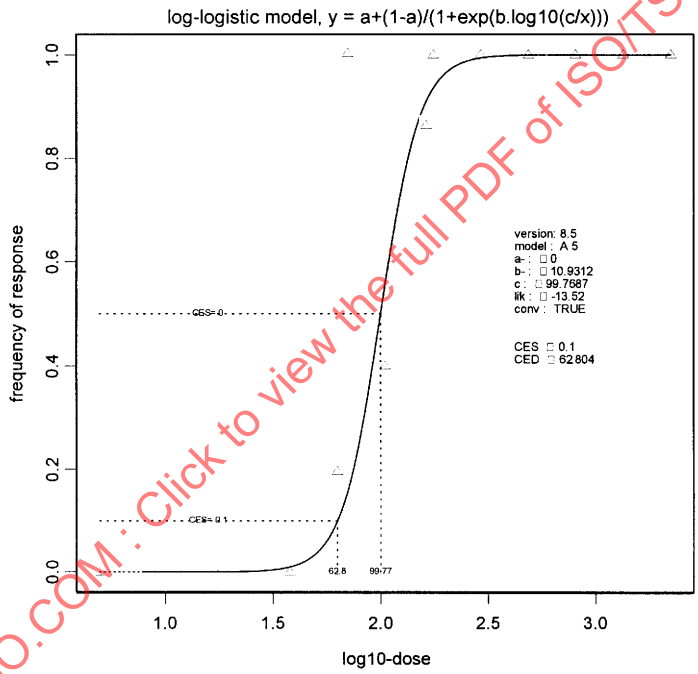
12) SAS, SPSS and splus are examples of a suitable products available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of these products.

13) The PROAST (Possible Risk Obtained from Animal Studies) software is available upon request (Wout.slob@rivm.nl).

As already discussed, it is theoretically impossible that the probability of response in the controls equals (strictly) zero. Therefore, the background response should be regarded as an unknown value, and be estimated from the data, even if the observed background response is zero (the fact that all observed control individuals did not respond does not imply that a response is impossible). Nonetheless, as Figure 10 illustrates, the background response may be estimated to be (virtually) zero, and in such situations fixing the background response at zero versus estimating it as a free parameter in the model does not make much difference (although the confidence intervals could be different, but probably not too much). Of course, one may always compare both ways of analyses in any practical situation.

It should be noted that omitting the background parameter from the model has the advantage of one less parameter to be estimated (parsimony principle), but at the same time the observations in the control group are made worthless in that way.

In practice, it may happen that the best fit of the model results in a negative estimate of the background response. To prevent this, the model should be fitted under the constraint that the background response shall be nonnegative (i.e. $a \geqslant 0$). Instead of a negative estimate, the background response a associated with the best fit is, in those situations, then estimated to be zero.



NOTE Here background mortality (parameter a) was included as a free parameter in the model, and estimated to be close to zero. The dashed lines indicate the LD_{50}, and the LD_{10}.

Figure 10 — Logit model fitted to mortality dose-response data (triangles), with background mortality

6.2.3.3 Analysis of data with various observed fractions at each dose group

Ecotoxicological (quantal) dose-response data often show replicated observed fractions at each concentration or dose group.

EXAMPLE 1 For example, the individual organisms in each dose group may be housed in different containers, each container resulting in an observed fraction of responding organisms.

EXAMPLE 2 As another example, the fraction of fertile eggs may be observed in individual female birds, where each dose group consists of various female birds.

In more general terms, these designs have various experimental units per dose group, and in each experimental unit the fraction of responding sampling units is counted. Of course, a dose-response model can

be fitted to such data by simply regarding the various observed fractions at each dose group as true replicates. In that case, it is assumed that the experimental units themselves (e.g. aquaria, of female birds) do not differ from each other.

If this cannot be assumed, the variability between experimental units shall be taken into account in the statistical analysis. Here, two approaches are briefly mentioned.

— One approach is to apply a normalizing (e.g. the square-root arcsine) transformation to the observed fractions related to each experimental unit. The transformed data can then be analysed as continuous data, as discussed in 6.3. However, this approach is problematic for data with 0 % and 100 % responses.

— Another approach is to account for the among-container variation by adjusting the binomial distribution. For example, the parameter reflecting the probability of response in the binomial distribution may be assumed to follow a beta distribution (reflecting the variability among containers). This implies that the observed response is beta-binomial distributed rather than binomial, and the associated likelihood may be maximized (see, for example, Teunis and Slob 1999).

Subclause 5.2.2.5. gives a description of two methods for deciding whether extra binomial variation is present.

6.2.3.4 Analysis of data with one observed fraction at each dose group

When the study design has only one container per dose group, the analysis appears at first sight simpler as compared to the situation of replicated containers at each dose. However, this is apparent only. If the containers differ by themselves, this between-container variation results in extra-binomial variation just as well. Theoretically, the variation among containers could be taken into account by the approaches mentioned above. However, experience with how this works in practical ecotoxicity data appears to be lacking.

6.2.3.5 Extrapolation and EC_x

Because of the fact that a fitted statistical model only reflects the information in the data, extrapolation outside the range of observation is usually unwarranted. Consequently, an EC_x that is estimated to be below the lowest applied (nonzero) dose should not be trusted.

6.2.3.6 Confidence intervals

Whatever definition for the EC_x is used, it is estimated from the point estimates of the parameters in the model. When these point estimates are obtained by maximum likelihood, these are Maximum Likelihood Estimates (MLEs). The EC_x is also a MLE when it is (indirectly) calculated from these values.

The MLE for the EC_{50}, or any other EC_x, is a point estimate only, and may, to a larger or smaller extent, be imprecise. The imprecision may be quantified by the standard error of the estimate, but it is more informative to calculate a confidence interval. A confidence interval indicates the plausible range for the parameter, e.g. a 95 %-confidence interval is supposed to contain the true value of the parameter with probability 95 %.

Confidence intervals may be assessed in various ways:

— plus or minus twice the parameter's standard error (provided by most dose-response software), which is estimated by the second derivative of the likelihood function (Hessian or information matrix), possibly with Fieller's correction (Fieller, 1954);

— based on the profile of the log-likelihood function, using the Chi-squared approximation of the log-likelihood;

— value(s);

— bootstrap methods (see e.g. Efron 1987, Efron and Tibshirani 1993);

— Bayesian methods, in particular if one has some preliminary knowledge on the plausible range of the parameter.

Various studies have compared the first three methods (see e.g. Moerbeek *et al*. 2004).

6.2.4 Assumptions

6.2.4.1 General

A dose-response model consists of a deterministic part (the predicted dose-response relationship) and a stochastic part (describing the noise). The assumptions made in the statistical part are analogous to those in hypothesis testing, and are only briefly mentioned here. The focus in this clause is on the additional assumption, that of the (deterministic) dose-response model.

6.2.4.2 Statistical assumptions

The assumptions for hypothesis testing equally hold for dose-response modelling:

— binomial distribution for observations per experimental unit, i.e. independence between the animals in the same experimental unit (e.g. container). When the experimental unit is not accounted for in the statistical model, it is additionally assumed that experimental units do not vary among each other by themselves (i.e. at the same dose);

— no systematic differences (caused by unintended experimental factors) between dose groups (the latter is particularly relevant for unreplicated designs, i.e. one container per dose-group);

— the values of the concentrations/doses are assumed to be known without error, or, in situations where they are measured, the measurement errors are assumed to be negligible.

Additional assumption:

— the fitted model has a shape that is close to the true dose-response relationship.

6.2.4.3 Evaluation of assumptions

6.2.4.3.1 Evaluation of basic assumptions

In designs with sample units (e.g. organisms, eggs) within experimental units (e.g. containers, female birds), the assumption of binomially distributed data may not be met, due to variation among the experimental units themselves. One way to check this is by fitting a model based on a betabinomial distribution, and comparing the associated log-likelihood with that obtained from a fit based on a binomial distribution. This comparison can be done by a likelihood ratio test, since the binomial and betabinomial distributions are nested.

6.2.4.3.2 Evaluation of the additional assumption

Fulfilment of the assumption that the shape of the fitted model is close to the true dose-response relationship depends not only on the choice of a proper dose-response model, but also on the quality of the dose-response data. Therefore, one not only needs to consider if the model is suitable to describe the data, but also if the data are good enough to sufficiently guide the model in obtaining the right shape. For a fuller discussion of evaluating the shape of the fitted dose-response model, see 6.4.

6.2.4.4 Consequences of violating the assumptions

6.2.4.4.1 Consequences of violating basic assumptions

When the assumption of binomial distribution is not met, due to variation between experimental units, a fitted quantal dose-response model may result in a biased estimate of the EC_x, as well as in too narrow confidence intervals.

6.2.4.4.2 Consequences of violating the additional assumption

Given that the data include both (close to) 0 % and (close to) 100 % responses, violation of the assumption (that the fitted model indeed reflects the true underlying dose-response relationship) is less serious for an EC_{50} than for an EC_x (the more so for lower values of x).

The (point) estimate of the EC_x may be inaccurate (biased), and the associated confidence interval may in extreme cases not even include the true value of the EC_x.

Therefore, it is not recommended to estimate an EC_x if the fitted model does not appear to be sufficiently confined by the data from visual inspection, or if it is found that various models fitting equally well result in different EC_x estimates. In the latter case, one might consider constructing an overall confidence interval for the EC_x based on various models that fit the data equally well (if repeating the experiment, aimed at more concentrations with partial responses, is not an option).

6.3 Dose-response modelling of continuous data (for a single exposure duration)

6.3.1 Purpose

While a quantal response is based on the observation of whether or not each single organism (biological system) has a particular property (e.g. death, clinical signs, immobilization), a continuous response is a quantitative measure of some biological property (e.g. body mass, concentration of enzyme). Such a continuous response is measured in each experimental unit. Since organisms (biological systems) are never identical by themselves or are not observed under identical conditions, the resulting data show a certain amount of scatter, depending on the homogeneity of the treatment group. This scatter may be assumed to follow a certain distribution, e.g. a normal, a log-normal, or a Poisson distribution.

Continuous data do not only differ from quantal data in a purely statistical sense (i.e. the underlying distribution). A more fundamental difference is that changes in response are interpreted in a completely different way. While the EC_x in quantal responses relates to a change in response rate, an EC_x in continuous responses relates to a change in the degree of the effect, as occurring in the average individual (of the population observed).

EXAMPLE An IC_{10} in a fish test is associated with a 10 % inhibition of the growth rate in the “average” fish (under the average experimental conditions).

The purpose of dose-response modelling of continuous data is to estimate the EC_x, where x is any given percentage. When the dose-response data relate to a single exposure particular duration, the estimated EC_x obviously only hold for that particular exposure duration (or to, e.g. a single acute oral dose).

6.3.2 Terms and notation

In this subclause, the following terms and notations are used. The continuous response, y, is related to the dose (or concentration), x, by the function, f:

$$y = f(x)$$

In ecotoxicology, the term EC_x is defined as the concentration (or dose) associated with an effect x[14], where x is defined as

$$x\,\% = 100\left(\frac{y(EC_x)}{y(0)} - 1\right)\%,$$

i.e. x is defined as a percent change in the (average) level of the endpoint considered, e.g. a 10 % decrease in mass.

In human toxicology, different terms exist for the EC_x. The equivalent terms are

— CED (Critical Effect Dose), which is equivalent to the EC_x, and

— CES (Critical Effect Size), which is equivalent to x in EC_x (see e.g. Slob and Pieters 1998).

However, in human toxicology another approach has been proposed, which is based on a change in response rather than on a change in the degree of effect. In that approach (also called the hybrid approach), the terms BMD and BMR are used (e.g. Crump 1995, Gaylor and Slikker 1990), but these terms are not comparable to the EC_x in continuous responses in ecotoxicology. The following table summarizes the terms.

Table 5 — Terms in human toxicology

	Ecotoxicology	Human toxicology term
Quantal response (x in terms of response)	x EC_x	BMR (benchmark response) BMD (benchmark dose)
Continuous response (x in terms of degree of effect)	x EC_x (IC_x)	CES (critical effect size) CED (critical effect dose)
Continuous response (BMR in terms of response)	— —	BMR BMD

6.3.3 Choice of model

6.3.3.1 First distinctions

A (statistical) dose-response model only serves to smooth the observed dose response, to estimate an EC_x by interpolating between applied doses, and to provide for a tool to assess confidence intervals. A statistical regression model itself does not have any meaning, and the choice of the model (mathematical expression) is largely arbitrary. Numerous dose-response models are theoretically possible, but in practice only a limited number are applied, mostly determined by historical habits in the field of application. A number of useful (families of) models are discussed here.

A first distinction that can be made is linear versus non-linear regression models. This distinction is made as the type of calculations is different between these two classes of models. In linear models, the calculations are relatively simple, and could be done without a computer, which is hardly possible for non-linear models. Clearly, given the widespread use of computers, this advantage has become more and more irrelevant, and non-linear models are gaining attention, as they may be considered to more realistic for reflecting a dose-response relationship (see below). Yet, linear models are briefly discussed, for the sake of completeness. After that, a number of other models (or family of models) is discussed, most of which are non-linear.

14) Note that x is used for both concentration (dose) and the degree of effect.

6.3.3.2 Linear models

Linear regression models are defined as models that are linear with respect to their parameters. They can be nonlinear with respect to the independent variable and thus not only include the straight line, but also quadratic, or higher order polynomials:

$$y = a + bx$$
$$y = a + bx + cx^2$$
$$y = a + bx + cx^2 + dx^3$$

etc.

These models have the property that the parameters (a, b, etc.) in the model can be estimated by evaluating a single (explicit) formula (as opposed to non-linear models, see below), which makes them relatively easy to apply.

Another advantage is that these models are nested. For example, the quadratic model can be turned into a linear model by taking $c = 0$. Inversely, a linear model can be turned into a quadratic model by incorporating an additional parameter (here: c). It can be statistically tested, if the addition of parameters leads to a significant improvement of the fit (e.g. by an F-test).

Linear models may be incorporated in the framework of GLM (generalized linear models), see e.g. Bailer and Oris (1997).

A disadvantage of linear models is that they are not necessarily strictly positive, while biological endpoints typically are (if the data are not pre-treated), which makes them theoretically implausible. Further, they are not necessarily monotone, which can result in doubtful results, especially in the situation of a limited number of dose groups.

6.3.3.3 Threshold models

A threshold model is a model that contains a parameter reflecting a dose-threshold, i.e. a dose below which the change in the endpoint is (mathematically) zero. In general, a threshold model is given by

$$y = a \quad \text{if } x < c;$$
$$y = a + f(x - c) \quad \text{if } x > c \tag{6}$$

where

c denotes the threshold concentration; and

$f(x)$ may be any function.

For example, in the (“hockey stick”) model.

$y = a$ if $x < c$

$y = a + b\,(x - c)$ if $x > c$, the response is linear above the threshold.

The threshold concentration could be called an EC_0, i.e. an EC_x with $x = 0$. At first sight, the threshold concentration appears attractive, as it avoids the discussion of what value of x in EC_x is ecologically relevant. However, various objections can be raised against the use of threshold models. One of them is that the (point) estimate of the threshold can be dependent on the dose-response relationship, i.e. the function that is chosen for $f(x)$ in Equation (6).

6.3.3.4 Additive versus multiplicative models

Strict continuous data (e.g. masses, concentrations) observed in toxicity studies usually have nonzero values in unexposed conditions, and the question then is to what extent the compound changes that level. Clearly, the compound interacts with that background level, by whatever biological mechanisms. This idea may be expressed in simple mathematical terms by incorporating the background level (a) in the dose-response model in a multiplicative way:

$$f(x) = a \cdot g(x) \tag{7}$$

rather than in an additive way:

$$f(x) = a + g(x) \tag{8}$$

as is more common in models discussed in statistical textbooks.

NOTE The models based on quantal models discussed in the previous subclause are also additive.

Of course, the whole idea of defining the EC_x as a given *percent* change compared to the background level is in concordance with the multiplicative interaction between compound and background level, as expressed in Equation (7). A further convenience of the multiplicative model is that two populations (e.g. species, sexes) showing different background levels but equally sensitive to the compound are, in this way, characterized by the same $g(x)$. This implies that in the multiplicative model, two equally sensitive populations (but possibly with different background levels) are defined to have the same EC_x.

6.3.3.5 Models based on "quantal" models

Continuous dose-response data from ecotoxicity tests have often been described by dose-response models that are derived from the models used for quantal data, i.e. models whose predicted values range from zero to one. To make these models applicable to continuous models, they are usually adjusted as follows:

$$y = y(0) + [y(\infty) - y(0)]\, f(x)$$

for increasing dose responses, and

$$y = y(\infty) + [y(0) - y(\infty)]\, [1 - f(x)]$$

for decreasing responses (see e.g. Bruce and Versteeg 1992, Scholze *et al.* 2001),

where

$y(0)$ is the (predicted) background value;

$y(\infty)$ is the (predicted) value at infinite dose; and

$f(x)$ is any quantal dose-response model.

NOTE These models are multiplicative (with respect to the background response), while their shape is typically sigmoidal.

EXAMPLE When the logit model is chosen for $f(x)$, the associated model for the continuous data becomes

$$y = y(0) + \frac{y(\infty) - y(0)}{1 + \exp\left[b \ln(ED_{50}/x)\right]}$$

In current practice, it is common to correct the data for the background response, and fit the model without a background parameter. As discussed in 4.3.5, this procedure of pre-treatment of the data ignores the estimation error in the observed background, and is therefore unsound. By incorporating the parameter $y(0)$ in

the model to be fitted, the estimation error is taken into account, and therefore this approach should always be taken.

6.3.3.6 Nested non-linear models

Slob (2002) proposed to use the following nested family of multiplicative non-linear models for general use in dose-response modelling.

Model 1:	$y = a$	with $a > 0$
Model 2:	$y = a \exp(x/b)$	with $a > 0$
Model 3:	$y = a \exp[\pm(x/b)^d]$	with $a > 0$, $b > 0$, $d \geqslant 1$
Model 4:	$y = a\,[c - (c - 1) \exp(-x/b)]$	with $a > 0$, $b > 0$, $c > 0$
Model 5:	$y = a\,\{c - (c - 1) \exp[-(x/b)^d]\}$	with $a > 0$, $b > 0$, $c > 0$, $d \geqslant 1$

where

y is any continuous endpoint; and

x denotes the dose (or concentration).

In all models, the parameter

a represents the level of the endpoint at dose zero, and

b can be considered as the parameter reflecting the efficacy of the chemical (or the sensitivity of the subject).

At high doses, Models 4 and 5 level off to the value ac, so the parameter

c can be interpreted as the maximum relative change.

Models 3 and 5 have the flexibility to mimic threshold-like responses (i.e. slowly changing at low doses, and more rapidly at higher doses).

All these models are nested to each other, except Models 3 and 4, which both have three parameters.

In all the models, the parameter a is constrained to being positive for obvious reasons (it denotes the value of the endpoint at dose zero). The parameter d is constrained to values larger than (or equal to) one, to prevent the slope of the function at dose zero being infinite, which seems biologically implausible. The parameter b is constrained to be positive in all models. Parameter c in Models 4 and 5 determines whether the function increases or decreases, by being larger or smaller than unity, respectively. To make Model 3 a decreasing function, a minus sign has to be inserted in the exponent.

These models have the following properties.

— The predicted response is strictly positive.

— They are monotone, i.e. either decreasing or increasing.

— They do not contain a threshold, but they are sufficiently flexible to show strong curvature at low doses, so as to mimic threshold-like responses.

— They can describe responses that level off at high doses.

— Two populations that differ in background level but are equally sensitive can be described by the same model, with only parameter a being different between the populations.

— It can be easily tested if two populations differ in sensitivity (by the likelihood ratio test).

— When two populations differing in sensitivity can be described by the same model from this family, with only parameter b (and possibly a) being different between the two populations, the difference in sensitivity can be quantified as the ratio of the value of b. This way of expressing differences in sensitivity is analogous to the relative potency factor, and to the extrapolation factors used in risk assessment.

For all five models, the EC_x can be derived by evaluating an explicit formula:

$$EC_x = \left\{ - \frac{\ln[\,(x + 1 - c) / (1 - c)\,]}{b} \right\}^{1/d}$$

where

x is defined as $x = y(EC_x)/a - 1$;

$c = 0$ for Models 2 and 3; and

$d = 1$ for Models 2 and 4.

Clearly, the five multiplicative models given here only apply for those endpoints that are strictly positive and have a nonzero background value (value of y in unexposed conditions). For example, describing internal concentration as a function of external concentration is not possible with these models, as in that case, y is expected to be zero for $x = 0$.

The procedure of selecting a model from this nested family of models, i.e. accepting additional parameters only when it results in a significantly better fit, is illustrated in Figure 11.

In this dataset, the following log-likelihoods were found:

— Model 1: 277,02;

— Model 2: 339,90;

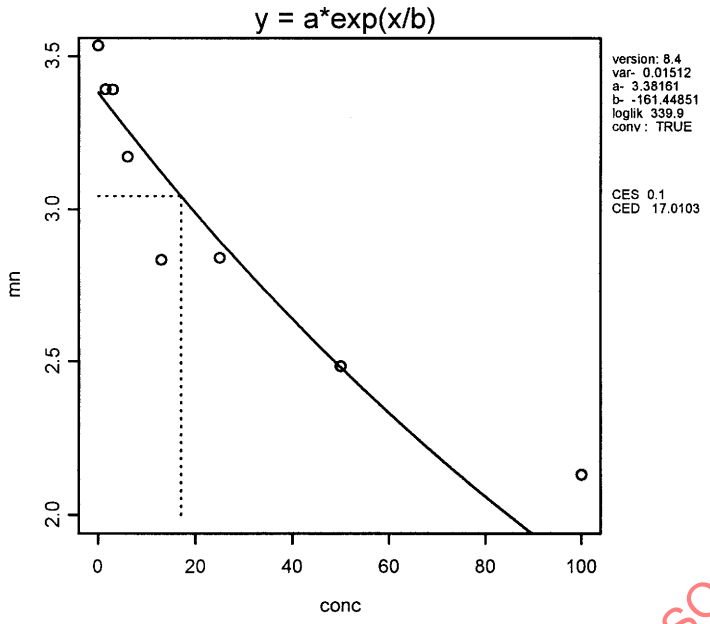
— Model 3: 339,90;

— Model 4: 351,11;

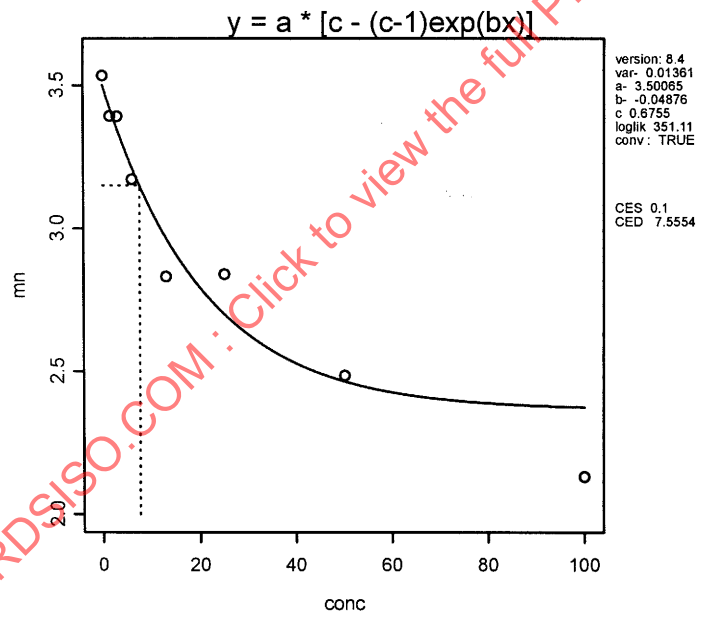
— Model 5: 351,11.

Model 3 resulted in exactly the same fit[15)] as Model 2, while Model 5 resulted in the same fit as Model 4. But Model 4 is significantly better than Model 2 (critical difference is 1,92 at $\alpha = 0{,}05$, according to the likelihood ratio test) and, therefore, Model 4 should be selected for this dataset. (Note that Model 3 and Model 4 are not nested; they both have three parameters).

15) Adding a parameter to a model can, by definition, not result in a lower (optimum) log-likelihood. When the log-likelihood remains the same, the additional parameter is estimated at the value that makes it disappear. In this case the parameter d was estimated to be one.



a) Exponential model



b) Improved exponential model, with additional parameter, c, enabling the response to level-off

NOTE The marks indicate the observed (geometric) means of the observations.

Figure 11 — Two members from a nested family of models fitted to the same data set

6.3.3.7 Hill model

Enzyme kinetics and receptor binding are usually described by the Hill model. It was introduced by A.V. Hill in 1910 in order to model the binding of oxygen to haemoglobin. The model is well known by enzymologists, biochemists and pharmacologists, and could be considered as one of the very few examples of a mechanistically based model. It has the form

$$y = \frac{a \quad x^c}{b \quad + \quad x^c}$$

where c is called the Hill parameter.

By setting $c = 1$, it is equivalent to the Michealis-Menten expression in a strict sense, with

a denoting the maximum level of y at infinite dose, and

b denoting the ED_{50} (dose resulting in half the maximum response).

The following formulation makes more sense for toxicology, since the parameter noted b in the Hill model is actually a thermodynamic equilibrium dissociation constant, K_d , that can be changed as EC_{50}^n which is more familiar to toxicologists and is homogenous to a concentration (or dose):

$$y = y(0) + \left[y(\infty) - y(0) \right] \frac{x^n}{EC_{50}^n + x^n}$$

It is worth noticing that the Hill model is analytically equivalent to the logit model:

$$y = \left[1 + e^{n \ln\left(\frac{EC_{50}}{x} \right)} \right]^{-1} = \left[1 + \left(\frac{EC_{50}}{x} \right)^n \right]^{-1} = \frac{x^n}{EC_{50}^n + x^n}$$

It should be noted that dose-response data observed in *in vivo* studies are not the result of a single underlying receptor binding process, but of many processes acting simultaneously. Yet, it may be a very accurate model for describing particular data, see e.g. Figure 19.

6.3.3.8 Non-monotone models

In some cases, dose-response data appear to be non-monotone. Unfortunately, it is not easy to assess if this is due to an underlying dose-response relationship that is indeed non-monotone. It is not unlikely that an apparent non-monotone dose response in observed data is due to experimental artefacts, either systematic errors in unreplicated dose groups or simply random noise. Although the latter possibility can be checked by statistical methods, the former cannot. Therefore, when the apparent monotonicity is based on a single treatment group, no unambiguous conclusion can be drawn. Only multiple dose studies with a clear non-monotone pattern, supported by various consecutive dose groups, may provide evidence of a real non-monotone response.

When it is assumed that the data do not contain any systematic errors, the straightforward way to test for non-monotonicity is by fitting a non-monotone model to the data, and by comparing the fit with a nested model that is monotone. For an example of a nested non-monotone model, see Brain and Cousens (1989) or Hoekstra (1993).

If the non-monotone model appears to be significantly better, it may still be doubtful if this particular model reflects the true dose-response relationship. The practical difficulty is that non-monotone models are highly data-demanding, in particular with respect to the number of consecutive dose groups around the local maximum (or minimum) of the response. Otherwise, the location and height of the local maximum response

are highly model-dependent. Therefore, fixation of the local maximum response requires the enclosure by sufficiently close adjacent dose groups. Since the location of the local maximum response is not known in advance, the study design would require a large number of dose groups. Therefore, when non-monotone dose-response relationships may be expected (as in plant growth data), a larger number of dose groups needs to be incorporated in the study design.

Of course, dose-response models include more parameters to be estimated, and this is another reason that many dose groups are required. In most practical data sets, various non-monotone models would give different results, and therefore can often not be trusted.

6.3.4 Model fitting and estimation of parameters

6.3.4.1 Software and assumptions

Fitting a model to dose-response data may be done by using any suitable software, e.g. SAS (www.sas.com), SPSS (www.spss.com), splus (www.insightful.com)[16)], and PROAST (Slob, 2003). The user does not need to be aware of the computational details, but some understanding of the basic principles in nonlinear regression is required to be able to interpret the results properly. These principles are discussed in 6.7.

Furthermore, the user should be aware of the assumptions underlying the fit algorithm. For continuous data, it is often assumed that the data follow a normal or a log-normal distribution. In the latter case, a log-transformation is used to make the data (more closely) normally distributed. When a normal distribution with homogenous variances is assumed (possibly after transformation), maximizing the likelihood or minimizing the Sum of Squares amounts to the same thing (see 4.3.6). When another distribution is assumed (e.g. a Poisson for counts), the model may be fitted by maximum likelihood, based on the particular distribution assumed. The parameter values produced by maximum likelihood are also called the Maximum Likelihood Estimates (MLEs).

6.3.4.2 Response in controls

In all the models discussed here, the background response is incorporated as a model parameter in the model. This parameter should be estimated from the data, before deriving the EC_x or IC_x. Pre-treatment of the data (dividing all responses by the mean background response) should be avoided (see also 4.3.4).

6.3.4.3 Fitting the model assuming normal variation

When the original data are assumed to be normally distributed with homogenous variances, the model may be fitted by either maximizing the log-likelihood function based on the normal distribution, or by minimizing the sum of squares. Both methods result in the same estimates of the regression parameters, which are maximum likelihood estimates (MLEs) in both cases. The fitted model describes the arithmetic mean response, as a function of dose.

6.3.4.4 Fitting the model assuming normal variation after log-transformation

When the residual variation is assumed to be log-normal, the model may be fitted after first log-transforming both the model predictions and the data, and then either maximizing the log-likelihood function based on the normal distribution, or minimizing the sum of squares. Both methods result in the same estimates of the regression parameters and the residual variance, which are maximum likelihood estimates (MLEs) in both cases. It should be noted that the resulting parameter estimates do relate to the original parameters of the (untransformed) model. Substituting the estimated regression parameters in the model results in a prediction of the median (or geometric mean) response as a function of dose. Therefore, in plotting the model together with the data, the back-transformed means (which are equivalent to the geometric means) should be plotted (see e.g. Figure 11).

16) SAS, SPSS and splus are examples of a suitable products available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of these products.

While the MLEs of the regression parameters relate to the model on the original scale, the MLE of the variance (s^2) relates to the log-transformed data. Apart from this variance (s^2) on log-scale, the variation of the scatter around the model (i.e. of the regression residuals) may be equivalently reported by the geometric standard deviation (GSD), which is the back-transformed square root of s^2, or by the coefficient of variation (CV), which relates to s^2 by

$$CV = \sqrt{\exp(s^2) - 1}$$

when s^2 relates to the variance of the data after natural log-transformation, or by

$$CV = \sqrt{\exp[s^2 \ln(10)] - 1}$$

when the $\log_{10}$-transformation was applied to the data.

At first sight, a disadvantage of taking the logarithm of the data before fitting is that the logarithm of zero does not exist. Although zero observations for continuous responses rarely occur in ecotoxicity testing, the following may be noted. Zero observations usually mean that the response is below the detection limit rather than truly zero. By regarding zero observations as truncated observations, they can be easily and accurately dealt with by incorporating the information that the observation is lower than the detection limit in the log-likelihood function.

6.3.4.5 Fitting the model assuming normal variation after other transformations

When another transformation is applied to the data, the same transformation should be applied to the model, before maximizing the likelihood (or minimizing the SS). Both the fitted model and the transformed data may be back-transformed before plotting. Again, the resulting plot relates to the predicted and observed *median* response, as a function of dose (assuming that the transformation made the scatter symmetrical).

6.3.4.6 No individual data available

In reported studies (published papers), individual observations are not always given. Instead, means and standard deviations (or standard errors of the mean) for each dose group are commonly reported. Since the mean and standard deviation are "sufficient" statistics for a sample from a normal distribution, a dose-response model can just as well be fitted based on these statistics without any loss of information (except possible outliers), by adjusting the log-likelihood function (Slob, 2002). In the case of an assumed log-normal distribution, sufficient statistics are provided by the geometric mean and the geometric standard deviation, or by the (arithmetic) mean and standard deviation on log-scale. These can be estimated from the reported mean and standard deviation (Slob, 2002). Figure 11 exemplifies a dose-response analysis applied to the reported means and standard deviations, without knowing the individual data, but taking the reported standard deviations into account.

6.3.4.7 Fitting the model using GLM

Since the log-likelihood function directly derives from the postulated distribution, one may theoretically assume any distribution, and apply maximum likelihood for fitting the model based on that assumption. For a number of distributions (the so-called exponential family of distributions), one may make use of the theory of Generalized Linear Models (GLM), and use existing software without deriving and programming one's own formulae. The GLM framework is also useful for analysing data with replicated concentration groups.

The Poisson distribution is a member of this exponential family, and the existing GLM software can be directly used. Thus, one may assume this distribution for the analysis of counts, and check if the distribution is reasonable.

The gamma distribution is another example of a distribution belonging to the exponential family. This distribution can be directly dealt with by the existing (GLM) software (e.g. in SAS, SPSS, splus). The gamma distribution is very similar to the log-normal distribution regarding its behaviour of describing the variation in

data. Therefore, an analysis based on either one of these two distributions may be expected to give very similar results. However, there are a few differences.

— An analysis based on the log-normal distribution results in a model describing the median response (as estimated by the geometric means).

— An analysis based on the gamma distribution describes the response in terms of the statistical expectation (as estimated by the arithmetic means).

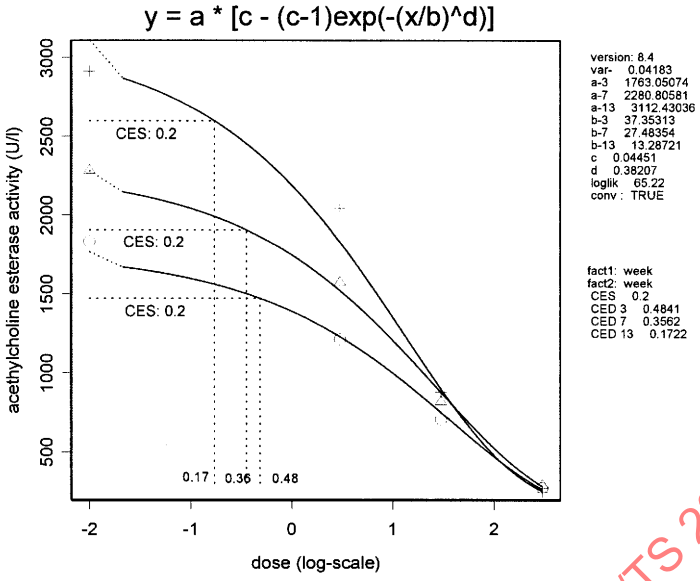
Therefore, the latter fitted model lies, on the whole, at a lower level than the former because the mean is larger than the median (the more so for larger experimental variation, i.e. more skewed scatter). However, both analyses may be expected to result in similar point estimates for the EC_x: the difference in level cancels as the EC_x is a ratio of the two medians, or of the two mean levels, respectively.

A second difference is, that the analysis based on the gamma distribution results in an estimate of the residual variation in terms of the variance on the original scale, while for the analysis based on the log-normal distribution, the residual variation is estimated in terms of a CV (or a GSD: geometric standard deviation).

6.3.4.8 Covariates

In many studies, not only the concentration is varied systematically. Other factors may also be varied intentionally as part of the design. For example, a chemical is studied under various conditions, e.g. temperature, pH or soil condition. Instead of fitting a model to each subset of data, it is often possible to fit the model simultaneously to the whole data set, by letting a particular parameter (possibly more) depend on that covariate.

Such an analysis is illustrated in Figure 12, where AChE inhibition was measured at three points in time, i.e. at three different exposure durations. Here, a four-parameter model was fitted, two of which were allowed to depend on duration. Thus a total of nine parameters was estimated, while a separate analysis for each duration would have resulted in a total of 15 estimated parameters (three times 4 regressions plus one variance parameter). The gain of this is that the resulting confidence intervals for the EC_x estimates are smaller.



Key

triangles 3 weeks

circles 7 weeks

pluses 13 weeks

NOTE Marks denote the geometric group means, the individual observations are not plotted here. The background AChE levels increase with duration (age), while the EC_x (CED for CES = 0,20) decreases with exposure duration. The model used is Model 5 from the nested family of models proposed by Slob (2002).

Figure 12 — Cholinesterase inhibition as a function of dose at three exposure durations

6.3.4.9 Heterogeneity and weighted analysis

In concordance with the parsimony principle (as discussed in e.g. 6.1), it is favourable to assume homogenous variances between dose groups; in this way, only one single parameter for the residual variance needs to be estimated. However, it should be noted that the term “homogenous variances” is closely associated with the normal distribution. When other distributions are assumed, the variances are generally not expected to be homogenous, e.g.:

— for log-normally (or Gamma) distributed data, variances increase with the means (more specifically, CVs are predicted to be constant), and this heterogeneity should vanish when the data are log-transformed. Thus, it may be assumed that (on the original scale) the CVs are homogenous, and the statistical analysis would result in a single estimate of the CV;

— for Poisson-distributed data (counts), the variances also increase with the mean. In fact they should be equal to the means, and if the data confirm this, no variance parameter needs to be estimated. In practice, this assumption is often violated, with the variances being larger than the means. This is called extra-Poisson variation, and an extra parameter may be estimated expressing the proportionality constant between mean and variance.

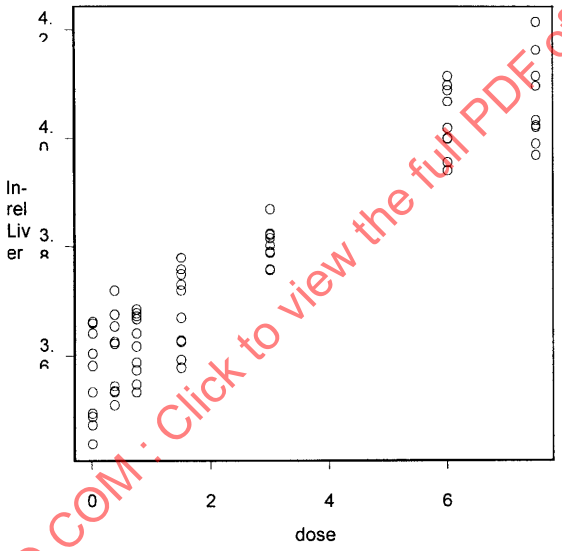
Apart from statistical reasons (the parsimony principle), the issue of homogenous variances should also be considered for biological reasons. It might be that the organisms did not respond equally to the compound due to variability in sensitivity, and this is reflected in the variances. It is not easy to discriminate between statistical heterogeneity (distribution effects) and biological heterogeneity (“true” effects). For that reason (among others), it is important to carefully consider what distribution should be assumed, e.g. by using historical data on the same (or similar) endpoint examined for other chemicals (or treatments).

When the heterogeneity of variances cannot be explained by the underlying distribution, one might conclude that the responses themselves are heterogeneous. Statistically, this implies that the precision of the estimated group means is not the same among groups. This may be taken into account in the statistical analysis by using a weighted analysis, e.g. weighted least squares, where the squares are multiplied by a weight, usually the inverse of the standard deviation of the relevant group, or by using maximum likelihood where a variance is estimated for each separate group[17]. For a more extensive discussion, see e.g. Scholze *et al*. (2001).

A weighted analysis should result in a more efficient estimate of the mean response (as a function of dose) in situations where the data are considered to reflect the same underlying response, and the heterogeneity is due to differences in measurement errors. The interpretation of a mean response is, however, problematic when the heterogeneity reflects that the population responds heterogeneously, in particular when this is caused by distinct subpopulations that differ in response.

EXAMPLE As an example, consider Figure 13, where relative liver masses are plotted on the log-scale (since for this endpoint, the scatter is normally proportional to the mean level). In this particular example, the scatter first decreases, then increases with the dose. This might lead one to perform a weighted analysis (e.g. weighted least squares).

However, as Figure 14 shows, the heterogeneity in variances is caused by different responses in males and females. Fitting the model taking sex into account results in two different dose-response relationships, each with homogenous scatter around it.



NOTE Normally, relative liver masses show homogenous scatter in log-scale, but in these data the scatter first decreases, then increases with dose.

Figure 13 — Relative liver masses against dose, plotted on log-scale

17) When the heterogeneity in response changes systematically with the dose in a way that cannot be explained by the underlying distribution, one may also incorporate a dose-response relationship for the variation parameter in the likelihood function.

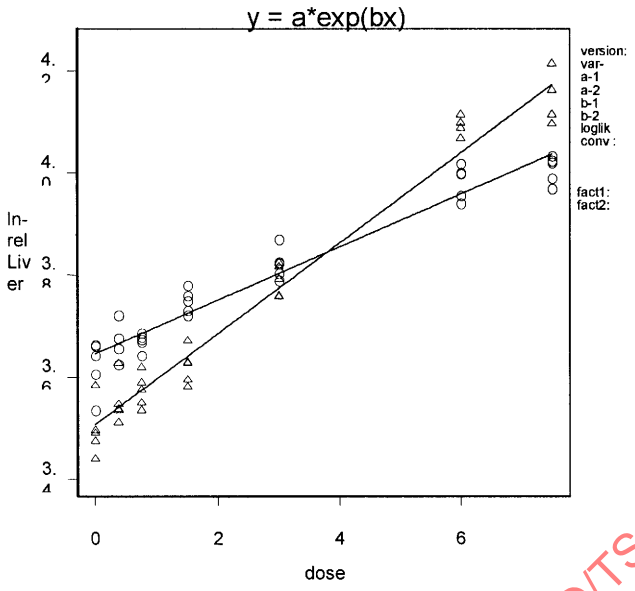


Figure 14 — Dose-response model fitted to the data of Figure 13, showing that the heterogeneous variance was caused by males (triangles) and females (circles) responding differently to the chemical

6.3.4.10 Confidence intervals

Confidence intervals may be assessed in various ways:

— the delta method, i.e. plus or minus twice (or the relevant standard normal deviate times) the standard error as estimated by the second derivative of the likelihood function (Hessian or information matrix); the standard errors of the parameters are provided by most dose-response software;

— based on the profile of the log-likelihood function, using the Chi-squared approximation of the log-likelihood;

— bootstrap methods (see e.g. Efron 1987, Efron and Tibshirani 1993);

— Bayesian methods, in particular if one has some preliminary knowledge on the plausible range of the parameter value(s).

The relative performance of the first three methods applied to a typical toxicological dataset (from a rodent study) has been examined by Moerbeek *et al*. (2004). In this study, the second and third method resulted in similar intervals, while the first method appeared less accurate.

6.3.4.11 Extrapolation

Because of the fact that a fitted statistical model only reflects the information in the data, extrapolation outside the range of observation is usually unwarranted. Therefore, estimating an EC_x that is much lower than the lowest applied (nonzero) dose or concentration should be avoided.

6.3.4.12 Analysis of data with replicated dose group

The individual organisms in each dose group may be housed in different containers. In that case, the individual observations may not be independent, due to systematic differences between the containers themselves. A straightforward and relatively simple approach for analysing such data is to follow two steps.

— In the first step the model is fitted as though the data were independent (i.e. the observations from various containers at the same dose are taken together and treated as a single sample).

— Then, the residuals from the fitted model are calculated and these are subjected to a nested analysis of variance, resulting in an estimate for the (residual) variance within as well as among the containers.

Strictly, the first step of this method is not completely valid, as it assumed independence between the observations. However, the results would normally not be much different (especially so for more or less balanced designs).

One may also fit a mixed model to the data, i.e. a model that contains both the (systematic) dose-response relationship and the random variation between containers.

These analyses result in an estimate of the variation among containers, and the residual variation within containers.

In studies without replicated dose groups, the variation between containers is incorporated into the residual variance. Theoretically, the variation between containers can still be estimated in designs with a sufficient number of dose groups, but practical experience with real toxicity data appears to be lacking.

6.3.5 Assumptions

6.3.5.1 General

A dose-response model consists of a deterministic part (the predicted dose-response relationship) and a statistical part (describing the noise). The assumptions made in the statistical part are analogous to those in hypothesis testing, and only be briefly mentioned here. The focus in this subclause is on the additional assumption, that of the (deterministic) dose-response model.

6.3.5.2 Statistical assumptions

The assumptions for hypothesis testing equally hold for dose-response modelling:

— independence between the animals in the same experimental unit (e.g. container);

— no variation between experimental units (e.g. containers) themselves, if they are not incorporated in the statistical analysis;

— a particular statistical distribution and variance structure for the residual variation, e.g.

  — normal distribution with homogenous variance;

  — log-normal distribution with homogenous Coefficient of Variation (CV);

  — gamma distribution with homogenous Coefficient of Variation (CV);

  — Poisson distribution without variance parameter, or with additional parameter for extra-Poisson variation;

— no systematic differences (due to unintended experimental factors) between dose groups;

— the values of the concentrations/doses are assumed to be known without error, or, in situations where they are measured, the measurement errors are assumed to be negligible.

6.3.5.3 Additional assumption

The shape of the fitted model is close to the true dose-response relationship.

6.3.6 Evaluation of assumptions

The statistical assumptions are similar to those in hypothesis testing, and may be further checked by plotting (analysing) the residuals (see 4.3.6 and Clause 5). However, the additional assumption (acceptance of the fitted dose-response model) is the most important, and the reader should first of all read and understand 6.4.

6.3.7 Consequences of violating the assumptions

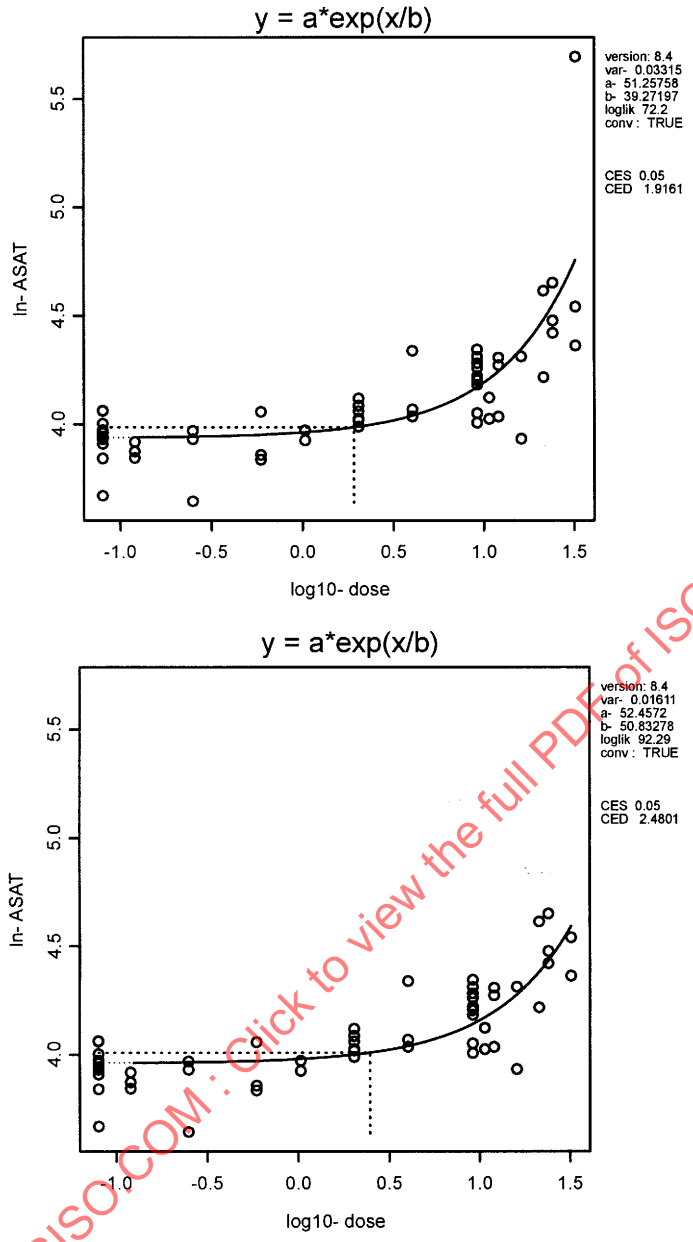
6.3.7.1 Basic assumptions

Violation of the assumption that containers do not vary amongst each other: While this variation is not taken into account in the statistical analysis, it does not have much impact on the point estimate of the EC_x (in particular when the number of replicates is similar between dose groups). It does, however, distort the estimate of the confidence interval, which is too narrow.

Systematic differences between (unreplicated) dose groups, caused by some unintended experimental factor, may have a deforming effect on the fitted model, and thereby result in a biased estimate of the EC_x. However, especially for multiple dose designs, the effect may be small: systematic deviations in particular dose groups are, to a greater or lesser extent (depending on the situation), mitigated by the other dose groups in the process of fitting a single dose-response model to the complete data set. To prevent systematic errors between dose groups as much as possible, attention should be paid to applying randomization procedures in the study protocol (see also 4.2.2).

If one suspects that experimental units (e.g. containers) vary by themselves, then one should incorporate replicated dose groups in the design (e.g. various containers per dose group), or increase the number of dose groups (keeping one container per dose). In both designs, the container effect can be estimated, although in the latter design this can only be done indirectly and may be difficult in practice.

A dose-response model is often relatively insensitive to outliers. See Figure 15 for an illustration.



NOTE The estimate of the EC_{05} (CED at CES = 0,05) is only mildly affected, even though the outlier is in the top dose. The 90 % confidence interval was estimated at (1,63 to 2,30) with the outlier included, and at (2,12 to 2,93) when excluded.

Figure 15 — Model fitted to dose-response data with and without an outlier in the top dose

6.3.7.2 Additional assumption

Violation of the assumption that the shape of the fitted model is close to the true dose-response relationship results in a biased estimate of the EC_x. There is no remedy against violation of this assumption, other than to repeat the study with an improved design. Therefore, it is not recommended to estimate an EC_x if the fitted model appears not sufficiently confined by the data from visual inspection, or if it is found that various models, fitting equally well, result in different EC_x estimates. In the latter case, one might consider constructing an overall confidence interval for the EC_x based on various models that fit the data equally well (if repeating the experiment, aimed at more concentrations with different response levels, is not an option).

6.4 To accept or not accept the fitted model?

6.4.1 Can the fitted model be accepted and used for its intended purpose?

A fundamental issue in dose-response modelling is the question

> "Can the fitted model be accepted and be used for its intended purpose (such as estimating an EC_x)?".

The issue is not that the model used should be the "right" model, since there is no such thing (at least not for statistical models). A statistical model completely hinges on the dose-response data, and the quality of the data is in fact the crucial aspect. In the fitting process, a model tries to hit the response at the observed doses. But when it is used for assessing an EC_x by interpolating between observed doses, the model should also "hit" the response in the non-observed dose range in between. In other words, there are two aspects in evaluating the fitted model: one should not only assess if the model succeeded in describing the observed responses, but also if the model can be trusted to describe the non-observed responses in between. The former aspect focuses on the quality of the model, the latter on the quality of the data. The following discussion indicates how to deal with these two aspects. It should be noticed that this discussion holds for both quantal and continuous dose-response data.

6.4.2 Is the model in agreement with the data?

This question may be addressed using the goodness of fit. Goodness-of-fit methods can be used in an absolute or in a relative sense. In an absolute sense, one may test if the data significantly deviate from a particular model. It should be noted that this test is sensitive not only to the inadequacy of the model chosen, but also to any violations of the basic assumptions (e.g. no independent observations, outliers). In particular, a single deviating concentration group (due to some unknown experimental factor) could make the model be rejected significantly even when it perfectly follows the overall trend in the data. Therefore, the (absolute) goodness-of-fit test should never be strictly applied. A visual check of the data is always needed and may overrule a goodness-of-fit test.

The goodness of fit may also be used in a relative way, i.e. to compare the fits of different models. When models are nested (as discussed in 6.2.2 and 6.3.2), the likelihood ratio test can be applied to determine the number of parameters needed for describing the data. For non-nested models, one may use the Akaike criteria (Akaike 1974, Bozdogan 1987), but this test is not exact.

It has been suggested to focus the goodness of fit to the region of interest (around the EC_x). This approach, in a sense, undermines the whole idea of dose-response modelling, i.e. describing the dose-response relationship as a whole. In particular, it is more sensitive for (systematic) errors in the data that happen to occur in one of the dose groups in the range of interest. As discussed in 6.3.4, one of the advantages of dose-response modelling is that potential systematic errors in a single dose group may be mitigated by the others.

6.4.3 Do the data provide sufficient information for fixing the model?

This question is at least as important as the previous. Therefore, the fitted dose-response model should always be visually inspected, not only to see if the data are close to the model, but also to check if the data provide sufficient information to confine the model. Here, one should ask the question

> "If additional data on intermediate dose groups had been available, could that have substantially changed the shape of the dose-response relationship as compared to the current fitted model?".

(See also 4.3.6.)

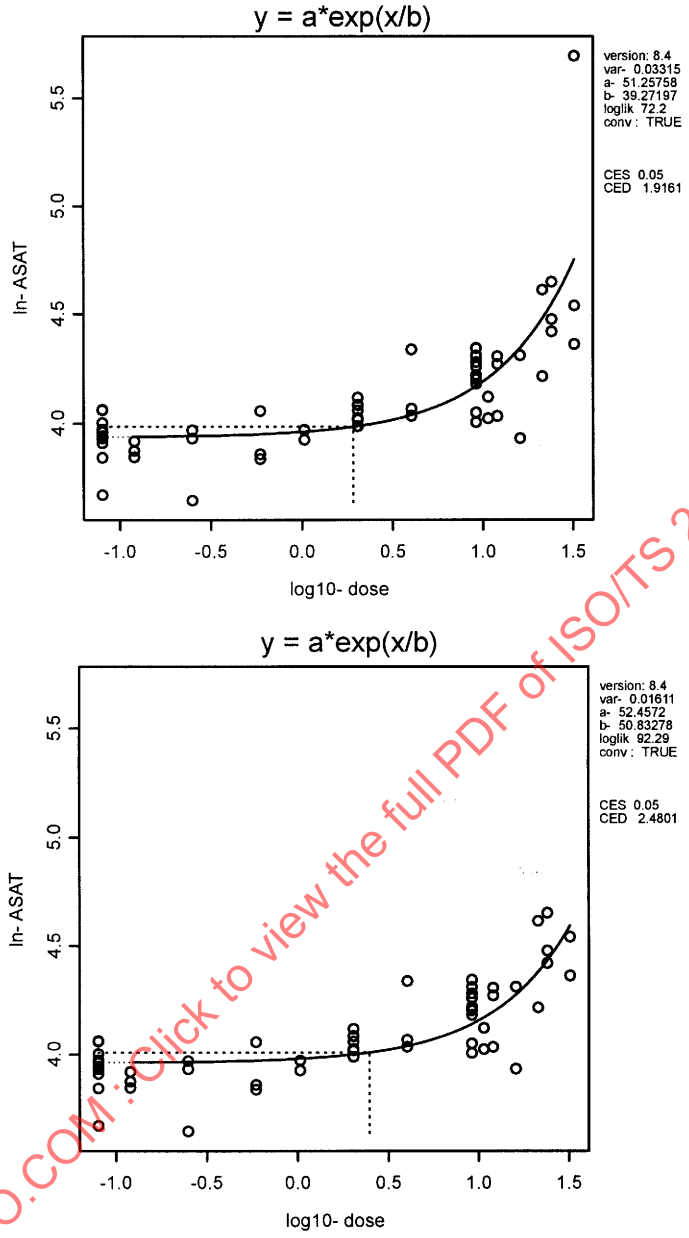
Another way to deal with this question is by comparing the outcomes from different fitted models. If the data do contain sufficient information to confine the shape of the dose-response relationship, different models fitting the data (nearly) equally well, result in similar fits and similar estimates of the parameters. To illustrate this (for the case of quantal data), the results of Figures 7 to 9 are summarized in Table 6. In this case, the results are quite independent from the model chosen, and one may conclude that the data provide sufficient information to rely on dose-response modelling.

Table 6 — Results of fitting three different models to the same data set (see Figures 7 to 9)

Model	a (background response)	LD_{50}	LD_{20}	Confidence interval of LD_{20} [a]	Log-likelihood
Probit	0,035 5	0,256 4	0,165	0,112 to 0,217	−34,01
Logit	0,035 6	0,255 4	0,167	0,121 to 0,220	−34,16
Weibull	0,035 2	0,262 5	0,145	0,084 to 0,218	−34,02

[a] Confidence intervals based on 1 000 parametric bootstrap runs.

As another illustration, Figure 16 shows two different models fitted to the same (continuous) data. Again, due to the good quality of the data, they result in very similar estimated concentration-response relationships, and therefore in a similar (point) estimate of any EC_x. In situations where the results (in particular, the EC_x) depend on the model chosen, it cannot be considered as a reliable estimate, and other methods should be considered (see 4.1).

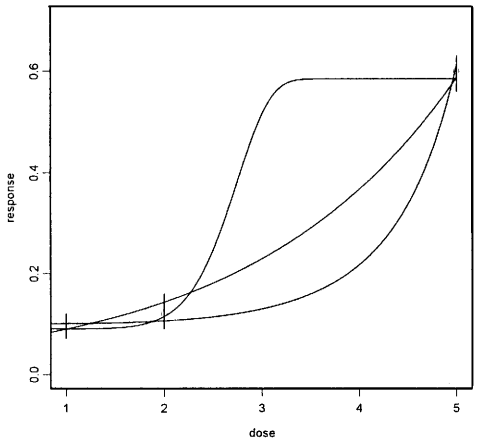


NOTE Small marks indicate individual observations; large marks indicate (geometric) means.

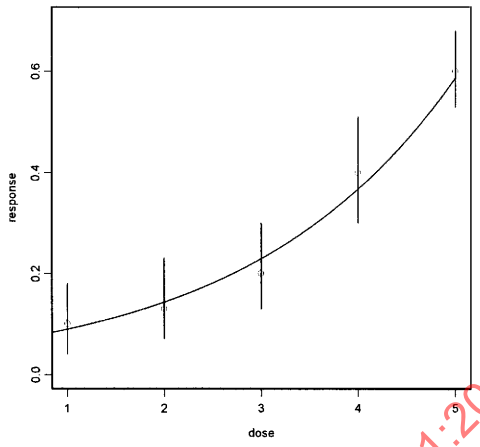
Figure 16 — Two different models (both with four parameters) fitted to the same data set resulting in similar dose-response relationships

In current practice, there is a tendency to focus on the first part and a formal goodness-of-fit test is often regarded as (sufficient) evidence that the model is acceptable. This is unfortunate, since a goodness-of-fit test tends to be more easily passed for data with few dose groups, and in exactly that situation, the second condition is more likely not to be met. In addition, a goodness-of-fit test assumes that the experiment was carried out perfectly, i.e. perfectly randomized with respect to all potentially relevant experimental factors and actions. Clearly, this assumption is not realistic.

The ideas discussed here are further illustrated (theoretically) in Figure 17. In the left panel, the data are insufficient to establish the dose-response relationship, leaving too much freedom to the model. In the right panel, the data are sufficiently informative to confine the shape of the dose-response relationship.



a) Data (either quantal or continuous) do not contain sufficient information to confine the dose-response relationship



b) Data (either quantal or continuous) do contain sufficient information to confine the dose-response relationship

NOTE 1 These figures also illustrate that more dose groups are more important than higher precision (indicated by vertical error bars). Although the precision of the EC_x estimate is lower in 17a), it is more likely to be biased.

NOTE 2 Dose group number 1, as indicated on the abscissa, may be read as the control group in these plots.

Figure 17 — Two data sets illustrating that passing a goodness-of-fit test is not sufficient for accepting the model

A number of general guidelines may be formulated in choosing and accepting a particular model for describing the dose-response data.

— When one of two nested models results in a significantly better fit, choose that model, otherwise the one with fewer parameters. One more parameter in the model can be regarded to result in a significantly better fit (at $\alpha = 0{,}05$) if the log-likelihood is increased by at least 1,92 (which is half the critical Chi-squared value with one degree of freedom at $\alpha = 0{,}05$). One may also follow this procedure as a proxy for non-nested models (or use Akaike's criteria).

— When two (or more) models have the same number of parameters, but one of them has a better goodness of fit, the choice of the better fitting model is obvious. However, if one prefers for some reason the other model, one may use Akaike's criteria to compare the model fits (Akaike, 1974; Bozdogan, 1987).

— When two models result in a similar goodness of fit, but their shapes are very different (resulting in different estimates of the EC_x), no conclusion can be made other than the data being inconclusive. In this situation it is not recommendable to derive an EC_x based on dose-response modelling.

— The situation that two (or more) models show a similar goodness of fit, both being similar in shape (resulting in similar EC_x estimates), this can be considered as a confirmation that the data provide sufficient information to assess the dose-response relationship, and estimate the EC_x.

It is re-emphasized that a dose-response model, as long as it is not based on the mechanism of action of the particular chemical, only serves to smooth the observed dose-response relationship, and to provide for a tool to assess confidence intervals. A statistical regression model itself does not have any biological meaning, and the choice of the model (expression) is to some extent arbitrary. It is the data, not the model, that should determine the dose-response relationship, and thereby the EC_x (Figure 17). When different models (with similar goodness of fit and equal number of parameters) result in different EC_x estimates, the data are apparently not suitable for dose-response modelling.

Dose-response models that are based on the mechanism of action of the particular chemical are, as opposed to statistical models, supposed to contain information by themselves, and therefore be less sensitive to data gaps (between dose groups). However, they do contain unknown parameters that need to be estimated from

the data, and it appears sensible to follow the guidelines described here in such models just as well. Mechanistic dose-response models are extremely rare, and contain some general elements at best. In the biological models discussed in Clause 7, the biological mechanisms in the models relate to the normal physiology in organisms rather than to the mechanism of action of specific chemicals.

6.5 Design issues

6.5.1 General

Concentration-response modelling can only be applied if the data contain sufficient information on the shape of the concentration-response relationship. Although this condition should be judged in each individual situation, experience teaches us that at least four different response levels are needed (including the control group) in the case of continuous data. A similar condition holds for quantal data, e.g. two partial kills next to (almost) complete mortality and (almost) complete survival. When one actually "knows" in advance that the concentration-response relationship is linear, designs with fewer concentration groups may be considered, and, as a matter of fact, they are more efficient in terms of precision. However, it seems rare that one can be confident *a priori* that the concentration-response is indeed linear (usually not much is known in advance about the tested chemical's action on the test organism) and extra concentration groups are highly recommendable.

A design with three concentration groups and a control may result in concentration-response data that allow for concentration-response modelling. However, it is always advisable to include more concentration groups for various reasons. If just one of the concentration groups was inadequately chosen (e.g. no observable response), concentration-response modelling fails. Further, systematic differences between treatment (concentration) groups are not unusual in toxicity testing (e.g. caused by systematic order in handling the groups), which may result in biased estimation of the concentration-response relationship. This unfavourable effect can be diminished in designs with more concentration groups.

In general, it may be stated that for the purpose of estimating an EC_x, it is important to have a sufficient number of dose groups, to prevent biased estimates of the EC_x. The allocation of the organisms (or experimental units) to more dose groups may be done at the expense of the number of replicates in each group without much loss (if any) for the precision of the estimated EC_x.

6.5.2 Location of dose groups

Concretely, for the purpose of estimating an EC_x, the available number of organisms (replicates) should be allocated to at least three (excluding the controls), but preferably more concentration groups. Next to a sufficient number of concentration groups (resulting in different response levels), one needs to choose a lowest and highest concentration level.

For quantal data, one may aim at four concentrations showing different response levels, including (nearly) none and (nearly) complete responses together with two concentrations with partial responses, as a minimum requirement. In continuous data, the low concentration is preferably chosen such that the observed response differs from the controls to a similar degree as x in the required EC_x (to prevent that the EC_x can only be estimated by extrapolation). Although one is usually interested in low response levels, high response levels are needed to assess the concentration-response relationship. The highest concentration would be preferably chosen such that the range between highest and lowest observed response is large enough to potentially include at least four different (in a rough statistical sense, that is, they appear detectable from the noise) response levels.

Interestingly, simulation studies show that the intuitive idea of concentrating dose levels around the EC_x is not optimal. Designs that include sufficiently high dose levels (or rather sufficiently different response levels compared to the controls) perform better (Slob, in preparation).

6.5.3 Number of replicates

In typical quantal data (with both 0 % and 100 % observed response levels), the precision of the EC_x declines with x, and the size of the experiment (total number of organisms or units) should be larger for smaller values

of x that are considered appropriate. Thus, when only an EC_{50} is required, a smaller experiment is required than when an EC_x is aimed for. In continuous dose response, this phenomena appears to be less prominent.

Theoretically, in quantal dose-response analysis, the relationship between the precision of an EC_x and the size of the experiment can be calculated. However, the number of organisms needed to obtain any given precision depends on the slope of the dose-response function itself, which is typically unknown before the study.

For the generally applicable nested family of (five) models, given in 6.3.2, simulation studies are being performed (for continuous data), to provide an indication of the (total) number of replicates necessary to achieve a particular precision for the EC_x (Slob, in preparation).

6.5.4 Balanced versus unbalanced designs

Due to the principle of leverage, observations in the extreme dose groups have more influence on the resulting model fit than the middle dose groups. This suggests that designs with larger sample sizes in the extreme dose groups may be more efficient than designs with the same sample sizes in all dose groups. Yet, preliminary simulation studies indicated that a design with twice the sample size in the controls performed only slightly better than one with equally sized dose groups. But more simulation studies are needed to give more definite answers to this question.

For designs with replicated experimental units (e.g. containers), where the number of replicates is small, say two, it appears wise to allocate a higher number of replicates in the controls, since a single erroneous replicate in the controls may then have a large impact on the model fit.

6.6 Exposure duration and time

6.6.1 General

It may be expected *a priori* that the response in biological systems is not only a function of dose, but also of the duration of the exposure. Therefore a model that describes the response as a function of both dose and duration would be more informative and give a more complete picture. Exposure duration is however a more complicated factor than dose, because it interferes with the factor time. The factor time has an impact by itself, e.g. on ageing, adaptation and repair of the processes underlying the response. Depending on the question to be answered, the study may, e.g.

— monitor the organisms during a period of time after an acute, or a fixed, period of exposure;

— monitor the organisms while they are held at various, but constant exposure levels;

— treat different groups of individuals with different exposure durations and compare the response at the end of exposure, or at a fixed point in time.

The second type of study is quite common in ecotoxicity testing in general (the others may be relevant for specific situations). Usually, in these studies the same (individual or groups of) organisms are followed over time. For example, the same organisms are recorded to have died or not. Or, egg production is monitored for the same (group of) organisms over time. In other studies, however, the observations in time may relate to different experimental units. As a result, the observations may be or not be independent, and this should be taken into account in the analysis of the data. This subclause briefly discusses the analysis of this type of data for both quantal and continuous data.

6.6.2 Quantal data

When a quantal response is observed at various points in time (e.g. number of additional deaths recorded each day while maintaining exposure at the same level), the statistical analysis of the dose-response data may be extended to include this extra information. Some authors have suggested fitting a dose-response model to the separate data sets, i.e. for each exposure duration separately, and plotting the ensuing EC_{50}s as a function of time. The value to which this function levels off is called the incipient EC_{50}, interpreted as the EC_{50} for “infinite” exposure. This is not a proper method and should be avoided, for various reasons.

— First, conceptual problems arise, e.g. an incipient LC_{50} does not make sense as more than 50 % of the animals die without any exposure at longer exposure durations.

— Second, statistical problems arise, e.g. the dose-response data at different time points are not independent, which hampers the establishment of confidence intervals for the incipient EC_{50}.

— And third, comparing dose-response models (such as the log-logistic) that are fitted for several time points separately may lead to inconsistent results (e.g. the fitted dose-response functions for various exposure durations intersect each other).

The approach of fitting a response surface to dose (concentration) and time simultaneously (multiple regression) is also improper, since the observations in time are not independent.

A proper way of modelling dose-time-response data where each individual is followed in time, is by assuming a relationship of dose with the hazard. The hazard[18)] reflects the probability of an individual to respond (e.g. die in the case of mortality) in a very small time interval, divided by the probability that it is still alive at that age. On a population level, this reflects the incidence of response during that small time interval, divided by the fraction of the population still alive at that age. By assuming that the hazard is a function of dose, the dose-time-response data can be described in a single model. The hazard can be directly transformed into a survival (or mortality) function, or, more generally, in a quantal time-response function. This function may be used for deriving the log-likelihood given the observed frequencies of response, in the usual way. There is a vast literature on survival analysis (see e.g. Cox and Oakes 1984, Miller 1981, Tableman and Kim 2004). For an example of dose-response modelling based on the hazard function, see 7.2.

6.6.3 Continuous data

6.6.3.1 General

For many continuous endpoints, observations can be (and sometimes are) made in time.

EXAMPLE 1 For example, body masses of animals can be determined at particular time intervals during the study. Or, the growth of algae can be monitored over time.

EXAMPLE 2 As another example, the number of eggs produced can be counted at specific time intervals.

It is re-emphasized that the observations in time may relate to the same or to different units (organisms), determining if the data should be treated as dependent or independent observations.

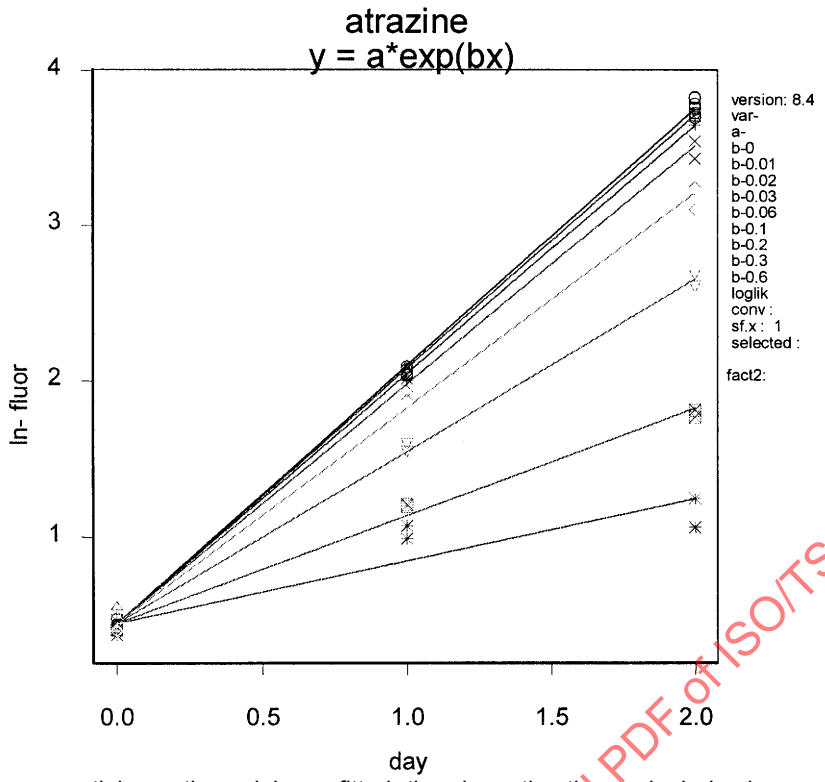
6.6.3.2 Independent observations in time

In some studies, the observations in time relate to different units.

EXAMPLE In algal growth studies, the biomass at a concentration is followed in time (e.g. day 0, 1 and 2). Suppose that once any of the algal test vessels has been measured, it is removed from the test. In that case, each observation relates to another vessel, and the data can be treated as independent, i.e. they can be taken together in a single analysis.

As an illustration consider the data in Figure 18, where at 9 different concentrations the biomass was measured at three consecutive days (each time with two replicates). Here a time-response model (i.e. a dose-response model with dose replaced by time) was fitted to all the data simultaneously, by assuming that the biomass at time zero was equal among the concentrations, while the growth rate differed between the concentrations. Thus, for each concentration a slope parameter b was estimated, but only a single parameter a and a single variance parameter. Thus, 11 parameters in total were estimated in a simultaneous fit of the model to these data.

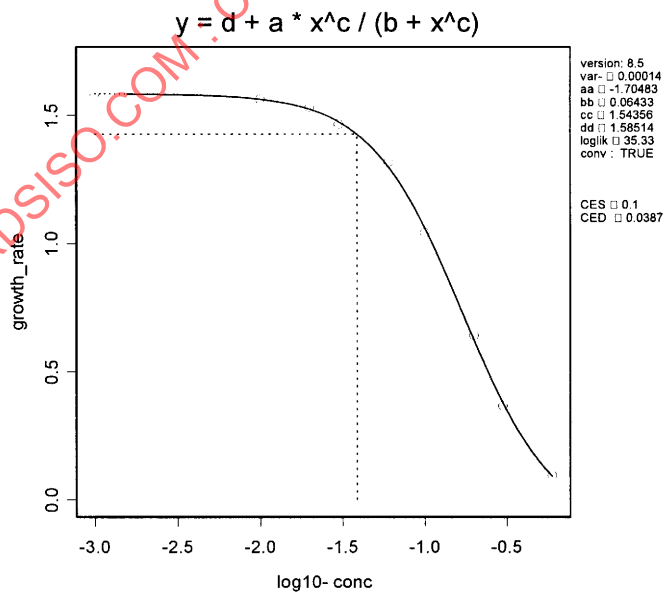
18) The hazard may be formally defined as $-\{[\mathrm{d}S(t)/\mathrm{d}t]/\ S(t)\}$, where $S(t)$ denotes the survival function.



NOTE Here, an exponential growth model was fitted, thereby estimating a single background value (a), a separate growth rate (b) for each concentration, and a single residual variance (for log-biomass). Note that replicates are treated as independent observations in this analysis.

Figure 18 — Observed biomasses (marks) as a function of time, for nine different concentrations of Atrazine

The estimated growth rates can subsequently be subjected to a dose-response analysis, as shown in Figure 19.

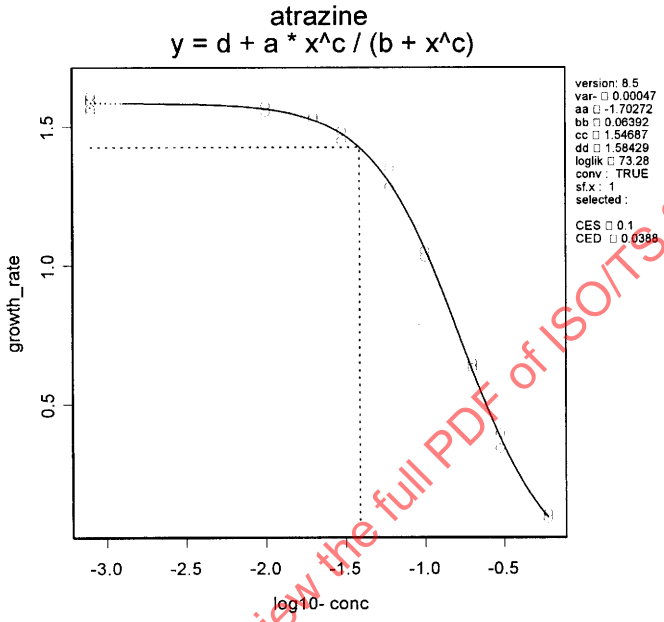


NOTE Point estimate of EC_{10} (= CED): 0,038 7 mg/l, with 90 % confidence interval: (0,035 1 to 0,042 4), based on 1 000 bootstrap runs.

Figure 19 — Growth rates as derived from biomasses observed in time (see Figure 18) at nine different concentrations (including zero), with the Hill model fitted to them

6.6.3.3 Dependent observations in time

When the data in time relate to the same experimental units, the observations cannot be treated as independent data, and an analysis as in Figure 18 is improper. When the data show a clear trend in time, a straightforward approach is to fit the exponential growth model to the biomasses, but now allowing each experimental unit (flask) to have its own growth rate. This amounts to fitting a separate time-response model for each separate experimental unit, and subsequently subject the relevant[19)] parameter estimates to a dose-response analysis. This analysis is analogous to that illustrated in Figures 18 and 19, but the concentration-response data now have replicates (see Figure 20).



NOTE Here, the individual flasks were taken into account, resulting in two growth rate estimates for each (nonzero) concentration, and six growth rates for concentration zero.

Point estimate of EC_{10} (= CED): 0,038 8 mg/l, with 90 % confidence interval: (0,035 5 to 0,042 1), based on 5 000 bootstrap runs.

Figure 20 — Estimated growth rates as a function of (log-)concentration Atrazine

It may be noted that the confidence intervals for the EC_{10} as derived from the data in Figure 19 and Figure 20 are very similar, despite the fact that in the latter case there are more data points. The reason is that the information in both data sets is in fact the same.

In data sets where no trend in time is apparent, one may just as well take the average over time (in each unit) and apply the dose-response analysis to the averages.

6.7 Search algorithms and non-linear regression

As discussed previously, non-linear regression models can only be fitted in an iterative “trial and error” approach. Computer software uses efficient algorithms to do that, and the user does not need to worry about the exact nature of the calculations. However, some basic understanding of the search process in required in order to interpret the results. In addition, such understanding is needed to evaluate whether the algorithm was successful or not, and if not, what if anything can be done about that.

19) The relevant parameter should follow from understanding the biological process. In algal biomass, the obvious parameter is the growth rate; when the observations relate to the number of eggs, supposed to level off at a constant level with age, either the parameter reflecting that level or the parameter reflecting the rate at which that level is reached could be the relevant parameter.

An iterative algorithm tries to find “better” parameter values in a process by evaluating if the fit can be improved by changing the parameter values. By regarding the fit criteria as a function of the parameters, the problem is in fact to find the maximum (in the case of likelihood) or minimum (in the case of Sum of Squares) of that function. Although algorithms have been developed to do this in an efficient way, one should keep in mind that the algorithm cannot see in advance where the optimum of the function is. One may compare the algorithm with a blindfolded person, who can only feel if there is a slope or not (and how steep it is). The algorithm recognizes the optimum by the property that around the optimum the slope changes from increasing to decreasing (or vice versa).

Obviously, the algorithm can only start searching when the parameters have values to start with. Although the software often gives a reasonable first guess for the starting values, the user may have to change these. It is not unusual (in particular when the information in the data is hardly sufficient to estimate the intended parameters) that the end result depends on the starting values chosen, and the user should be aware of that.

The algorithm keeps on varying the parameter values until it decides to stop. There are two possible reasons for the algorithm to stop the searching process.

— The algorithm has *converged*, i.e. it has found a clear maximum in the log-likelihood function. In this case the associated parameter values can be considered as the “best” estimates (MLEs if the likelihood was maximized). However, it can happen that the log-likelihood function has not one but more (local) maxima. This means that one may get other results when running the algorithm again, but with other start values. This can be understood by remembering that the algorithm can only “feel” the slope locally, so that it usually finds the optimum that is closest to the starting point.

— The algorithm has not converged, i.e. the algorithm was not able to find a clear optimum in the likelihood function, but it stops because the maximum number of iterations (trials) is exceeded. This may occur when the starting values were poorly chosen, such that the associated model would be too far away from the data. Another reason could be that the information in the data is poor relative to the number of parameters to be estimated. For example, a concentration-response model with five unknown parameters cannot be estimated with a four-concentration-group study. As another example, the variation between the observations within concentration groups may be large compared to the overall change in the concentration-response. In these cases, the likelihood function may be very flat, and the algorithm cannot find a point where the function changes between increasing and decreasing. The user may recognize such situations by high correlations between parameter estimates, i.e. changing the value of one parameter may be compensated by another, leaving the model prediction practically unchanged.

6.8 Reporting statistics

6.8.1 Quantal data

With quantal data, the following should be reported:

— test endpoint assessed;

— number of test groups;

— number of subgroups within each group (if applicable);

— identification of the experimental unit;

— nominal and measured concentrations (if available) for each test group;

— number exposed in each treatment group (or subgroup if appropriate);

— number affected in each treatment group (or subgroup if appropriate);

— proportion affected in each treatment group (or subgroup if appropriate);

— the dose metric used;

— the model function chosen for deriving the EC_{50} (EC_x);

— plot of dose-response data with fitted model, including the point estimates of the model parameters and the log-likelihood (or residual SS);

— fit criteria for other fitted models;

— the EC_{50} together with its 90 %-confidence interval;

— if required: the EC_x together with its 90 %-confidence interval;

— method used for deriving confidence intervals.

6.8.2 Continuous data

With continuous data, the following data should be reported:

— test endpoint assessed;

— number of test groups;

— number of subgroups within each group (if applicable);

— identification of the experimental unit;

— nominal and measured concentrations (if available) for each test group;

— the dose metric used;

— number exposed in each treatment group (or subgroup if appropriate);

— arithmetic group means and standard deviations, but geometric group means and standard deviation if log-normality was assumed;

— the model function chosen for deriving the EC_x;

— plot of dose-response data with fitted model, including the point estimates of the model parameters and the log-likelihood (or residual SS);

— fit criteria for other fitted models;

— the EC_x (CED) together with its 90 %-confidence interval;

— method used for deriving confidence intervals.

7 Biology-based methods

7.1 Introduction

7.1.1 Effects as functions of concentration and exposure time

Biology-based methods not only aim to describe observed effects, but also to understand them in terms of underlying processes such as toxico-kinetics, mortality, feeding, growth and reproduction (Kooijman 1997). This focus on dynamic aspects allows exposure time to be treated explicitly.

This clause focuses on the analysis of data from a number of standardized toxicity tests on mortality, body growth (e.g. fish), reproduction (e.g. daphnia), steady-state population growth (of e.g. algae, duckweed). The guidelines for these tests prescribe that background mortality is small, while the duration of the test is short

relative to the life-span of the test-organisms. Moreover the tests are done under conditions that are otherwise optimal, which excludes multiple stressors [e.g. effects of food restriction, temperature (Heugens 2001, 2003)], and quite a few processes that are active under field conditions [e.g. adaptation, population dynamics, species interactions, life-cycle phenomena (Sibly and Calow 1989)]. The type of data that are routinely collected in these tests are very much limited, and do not include internal concentrations of test compounds. These restrictions exclude the application of quite a few potentially useful methods and models for data analysis, such as more advanced pharmacokinetic models and time series analysis, see e.g. Newman (1995). The theory behind biology-based methods can deal with dynamic environments (changing concentrations of test compounds, changing food densities), but the application in the analysis of results from toxicity tests is simplified by the assumption that organisms' local environment in the test is constant.

Biology-based methods make use of prior knowledge about the chemistry and biology behind the observed effects. This knowledge is used to specify a response *surface*, i.e. the effects as a function of the (constant) concentration of test compound in the medium *and* the exposure time to the test compound. This response surface is determined by a number of parameters. The first step is to estimate these parameters from data. The second step is to use these parameter values to calculate quantities of interest, such as the EC_x-time curve, or the confidence interval of the No Effect Concentration (NEC). It is also possible to use these parameter values to predict effects at longer exposure times, or effects when the concentration in the medium is not constant. If the observed effects include those on survival and reproduction of individuals, these parameters can also be used to predict effects on growing populations (in the field) [Kooijman (1985, 1988, 1997), Hallam *et al.* 1989].

It is essential to realize that EC_x values decrease for increasing exposure time, as long as the exposure concentration and the organism's sensitivity remain constant. This is partly due to the fact that effects depend on internal concentrations (Kooijman 1981, Gerritsen 1997, Péry *et al.* 2002), and that it takes time for the compound to penetrate the body of test organisms. (The standard is to start with organisms that were not previously exposed to the compound.) The exposure period during which the decrease is substantial depends on the properties of the test compound and of the organism and the type of effect.

For test compounds with large octanol-water partition coefficients and test organisms with large body sizes, this period is usually large. The LC_{50} for daphnia hardly decreases for a surfactant after two days, for instance, but their LC_{50} for cadmium still decreases substantially after three weeks. For this reason, biology-based methods fit a response *surface* to data, using all observation times simultaneously. If just a single observation time is available, however, these methods can still be used and the response surface reduces to a response curve. Obviously, such data hardly contain information about the dynamic aspect of the occurrence of effects. The parameter(s) that quantify this aspect are then likely to be poorly defined. This does not need to be problematic for all applications (such as the interpolation of responses for other concentrations at that particular observation time; this is the job of dose-response methods).

It is strongly recommended, however, for a two-day test on survival, for instance, to use not only the counts at the end of the experiment, but also those at one day. Such data are usually available (and GLP even requires the reporting of those data), but these data are not always used. More recommendations are given in 7.8.

In practice, it is not unusual that very few, if any, concentrations exist with partial effects; survival of a cohort of individuals tends to be of the "all or nothing" type in most concentrations. High concentrations run out of surviving individuals more rapidly than lower concentrations. This can occur in ways such that for each single observation time, no, or very few, concentrations show partial mortality. This situation also occurs if each individual is exposed separately, and is measured rather than using nominal concentrations in the data analysis; one then has just a single individual per concentration because no two concentrations are exactly equal. Although such a case is generally problematic for dose-response methods, because a free slope parameter has to be estimated (Kooijman 1983), biology-based methods do not suffer from this problem, because the (maximum) slope is not a free parameter (model's slope of concentration-survival curves increases during exposure), and the information of the complete response surface is used. An example is given in 7.3.

Biology-based methods allow the use of several data sets simultaneously, such as survival data, sub-lethal effect data, and data on the concentration of test compound inside the bodies of the test organisms during accumulation/elimination experiments. As is discussed below, logical relationships exist between those data, and these relationships can be used to acquire information about the value of particular parameters that occur in all these data sets. Both the statistical procedures and the computations can become somewhat more

complex in this type of advanced applications, but free and downloadable software exist that can do all computations with minimum effort (see below).

7.1.2 Parameter estimation

The maximum likelihood (ML) method is used to estimate parameter values. (The criterion of least-squared deviations between data and model predictions is a special case of the ML method, where the scatter is independently normally distributed with a constant variance.) If more than one data set is used (for instance, data on body size and reproduction rate and/or internal concentration), the assumption is that the stochastic deviations from the mean are independent for the different data sets. This allows the formulation of a composite likelihood function that contains all parameters for all models that are used to describe the available data sets. For effects on survival, the number of dead individuals between subsequent observation times follows a multinomial distribution (see e.g. Morgan 1992); for sub-lethal effects, the deviations from the mean are assumed to be independently normally distributed with a common (data-set-specific) variance. The deterministic part of the model prediction is fully specified by the theory. For the stochastic part, only these straightforward assumptions are programmed in the DEBtox software (see 7.9.2). The software package DEBtool[20)], allows more flexibility in the stochastic model, e.g. for ML estimates in the case that the variance is proportional to the squared mean; this rarely results in substantially different estimates, however.

If surviving individuals are counted in a toxicity test, and tissue-concentrations are measured in another test, a composite likelihood function can be constructed that combines these multinomial and normal distributions. The elimination rate (dimension: per time) is a parameter that occurs in both types of data. In survival data, it quantifies how long it takes for death to show up; if the elimination rate is high, one only has to wait a short time to see the ultimate effects. The elimination rate can, therefore, be extracted from survival data in absence of data on internal concentrations. Although it is helpful to have the concentration-in-tissue data (both for estimating the parameters and for testing model assumptions), these data are by no means required to analyse effects on survival. If one has prior knowledge about the value of the elimination rate, one can fix this parameter and estimate the other parameters (such as the NEC) from survival data.

Profile likelihood functions are used to obtain confidence intervals for parameters of special interest, and in particular for the NEC. This way of quantification of the uncertainty in a parameter value does not necessarily lead to a single compact interval, but sometimes leads to two, non-overlapping intervals. Therefore, they can better be indicated with the term “confidence set”. Computer simulation studies have shown that these confidence sets are valid for extremely low numbers of concentrations and of test organisms (Andersen *et al.*, 2000).

Estimation procedures have been worked out (Kooijman 1983) to handle somewhat more complex experimental designs, in which living individuals are sacrificed for tissue analysis during the test. The information that they were still living at the moment of sampling is taken into account in the estimation of parameter values that quantify the toxicity of the compound. Péry *et al.* (2001) discuss the estimation of parameters in the case that the concentration in the media varies in time using hazard models; Kooijman (1981) and Reinert *et al.* (2002) use critical body residue models.

7.1.3 Outlook

This Technical Specification only discusses the simplest experimental designs of toxicity tests and the simplest models. The authors of this Technical Specification are unaware of alternative models in the open literature that are applicable on a routine basis, and hope that this Technical Specification stimulates research in this direction. The models can be, and have been, extended in many different ways. All individuals are assumed to have identical parameter values in the models that are discussed below. Individuals can differ, despite the standardization efforts in tests. Such differences might relate to differences in one or more parameter values (Sprague 1995). It is mathematically not difficult to include such differences in the analysis, on the basis of assumptions about the simultaneous scatter distribution of the parameter values. Needless to say, one really does know little if anything about this distribution. This makes such assumptions inspired by convenience arguments rather than by mechanistic insight. A strong argument for refraining from such

20) DEBtool and DEBtox are examples of suitable products available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of this/these products.

extensions is that the method becomes highly unpractical. The data simply do not allow a substantial increase in the number of parameters that shall be estimated from routine data.

The theory covers many features, such as extrapolating from constant to pulse exposures and vice versa, and including the effects of senescence, that are not yet worked out in software support (see 7.9).

7.2 Modules of effect-models

7.2.1 General

Effects are described on the basis of a sequence of three steps (modules):

a) **change in the internal concentration**: the step from a concentration in the local environment (here the medium that is used in the test) to the concentration in the test organism;

b) **change in a physiological target parameter**: the step from a concentration in the test organism to a change in a target parameter, such as the hazard rate, the (maximum) assimilation rate, the specific maintenance rate, the energy costs per offspring, etc.;

c) **change in an endpoint**: the step from a change in a target parameter to a change in an endpoint, such as the reproduction rate, the total number of offspring during an exposure period, etc.

This decomposition of the description of effects into three modules calls for an eco-physiological model of the test organism that reveals all possible physiological targets. The primary interest is in small effects. A simplifying assumption is that just a single physiological process is affected at low concentrations and that this effect can be described by a single parameter. At higher concentrations, more processes might be affected simultaneously. This means that the number of possible effects (and so the number of required parameters) can rapidly increase for large effects. It is unpractical and, for our purpose not necessary, to try to describe large effects in detail.

The concept “most sensitive physiological process” has an intimate link with the concept “no effect concentration”. The general idea is that each physiological process has its own “no effect concentration”, and that these concentrations can be ordered. Below the lowest no effect concentration, the compound has no effect on the organism as a whole. Between the lowest and the second lowest no effect concentrations, a single physiological process is affected; between the second and the third lowest no effect concentrations, two processes are affected, etc.

The concept “**no effect concentration**” is quite natural in eco-physiology (see e.g. Chen and Selleck, 1969). All methods for the analysis of toxicity data (including hypothesis-testing and dose-response methods) make use of the *concept* “no effect concentration”. All methods assume, at least implicitly, that compounds in the medium, apart from the tested chemical, do not affect the organism’s response. Hypothesis testing explicitly assumes that the tested chemical has no effect on the response at concentrations equal to, and lower than, the NOEC. Biology-based methods use the NEC as a free *parameter*.

Generally each compound has three domains in concentration:

1) effects due to **shortage**. Think, for instance, of elemental copper, which is required in trace amounts for several co-enzymes of most species;

2) **no effect** range. The physiological performance of the organism seems to be independent of the concentration, provided that it remains in the no effect range. Think, for instance, of the concentration of nitrate in phosphate-limited algal populations; Liebig’s famous minimum law rests on the “no effect” concept (von Liebig 1840);

3) **toxic** effects. Think, for instance, of glucose, which is a nutritious substrate for most bacteria in low concentrations, but inhibits growth if the concentration is as high as in jam.

It is essential to realize that the judgement “no effect” is specific for the level of organization under consideration. At the molecular level, molecules cannot be classified into one type that does not give effects, and another type that gives effects. The response of the individual as a whole is involved (Elsasser, 1998).

EXAMPLE The concept “no effect concentration” can deal with the situation that it is possible to remove a kidney, for instance, from a human subject (so a clear effect at the suborganism-level), without any obvious adverse effects at the level of the individual (during the limited time of a test). This example, therefore, shows that below the NEC effects can occur at the suborganismic level (e.g. enzyme induction), as well as on other endpoints that are not included in the analysis (e.g. changes in behaviour).

Most compounds are not required for the organisms’ physiology, which means that their range of concentrations that cause effects due to shortage is zero, and the *lower* bound of the no effect range is, therefore, zero as well. Some compounds, and especially the genotoxic ones (van der Hoeven *et al.* 1990, de Raat *et al.* 1985, 1987, Purchase and Auton 1995), are likely to have a no effect range of zero as well, and the *upper* bound of the no effect range is, therefore, also zero. This gives no theoretical problems in biology-based methods. An NEC of zero is just a special case, and a point estimate for this concentration from effect-data should (ideally) not deviate significantly from zero (apart from the Type I error; a Type I error occurs if the null hypothesis is rejected, while it is true).

The model for each of the three modules for the description of effects is kept as simple as possible for practical reasons, where one usually has very little, if any, information about internal concentrations, or physiological responses of the test organisms. Each of these modules can be replaced by more realistic (and more complex) modules if adequate information is available. Some applications allow further simplification. Algal cells, for instance, are so small that the intracellular concentration can be safely assumed to be in instantaneous equilibrium with the concentration in the media that are used in the test for growth inhibition. This gives a constant ratio between the internal and external concentrations, and simplifies the model considerably. The standard modules are introduced below.

7.2.2 Toxico-kinetic model

The toxico-kinetic module is taken to be a first order kinetics by default; the accumulation flux is proportional to the concentration in the local environment, and the elimination flux is proportional to the concentration inside the organism. This simple two-parameter model is rarely accurate in detail, but frequently captures the main features of toxico-kinetics (Harding and Vass 1979, Kimerle *et al*. 1981, McLeese *et al.* 1979, Spacie and Hamelink 1979, Wong *et al.* 1981, Janssen *et al.* 1991, Legierse *et al*. 1998, Jager 2003, Jager *et al.* 2003). It can be replaced by a more-compartment model, or a pharmacokinetic model, if there are sound reasons for this. Metabolic transformation, and satiation in the elimination rate can modify toxico-kinetics in ways that are sometimes simple to model (Kooijman 2000).

If the organism grows during exposure, or changes in lipid content occur (for instance when the test organisms are starved during exposure), predictable deviations from first order kinetics can be expected, and taken into account (Kooijman and van Haren 1990, Kooijman 2000). Dilution by growth should always be taken into account in the test for body growth and reproduction, since such a dilution affects the effect-time profiles substantially.

7.2.3 Physiological targets of toxicants

The specification of sub-lethal effects involves an eco-physiological model that reveals all potential target parameters, and allows the evaluation of the endpoints of interest. A popular endpoint is, for instance, the cumulative number of offspring of female daphnia in a three-week period. The model should specify such a number, as well as the various physiological routes that lead to a change of this number. It should also be not too complex for practical application. An example of such a model is the Dynamic Energy Budget (DEB) model. Because it is the only model for which generic applications in the analysis of toxicity data has been worked out presently, the following discussion focuses on this model.

The DEB model results from a theory that is described conceptually in Kooijman (2001) and Nisbet *et al*. (2000), and discussed in detail in Kooijman (2000). Figure 21 gives a scheme of fluxes of material through an animal, which are specified mathematically in the DEB model, on the basis of mechanistic assumptions. The

model's main features are indicated in the legend of Figure 21. The DEB theory is not confined to animals, however, and covers all forms of life.

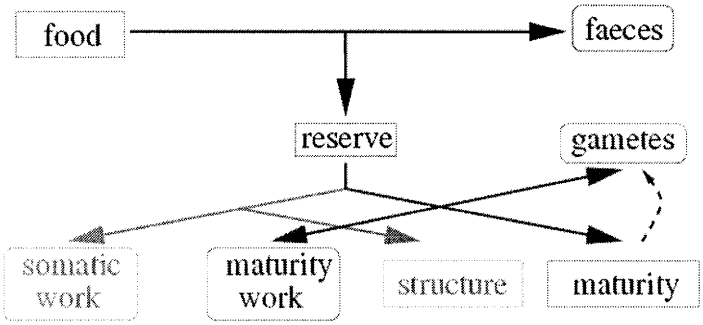


Figure 21 — Fluxes of material and energy through an animal, as specified in the DEB model

Assimilation, i.e. the conversion of food into reserve (plus faeces) is proportional to structure's surface area. Somatic and maturity work (involved in maintenance) are linked to structure's mass, but some components (heating in birds and mammals, osmo-regulation in freshwater organisms) are linked to structure's surface area. Allocation to structure is known as growth; allocation to maturity as development; allocation to gametes as reproduction. Embryos do not feed; juveniles do not reproduce; adults do not develop. Reserves and structure are both conceived as mixtures of mainly proteins, carbohydrates and lipids; they can differ in composition. The rate of use of reserve depends on the amount of reserve and structure; this rate is known as the catabolic rate. A fixed fraction of the catabolic flux is allocated to somatic maintenance plus growth, as opposed to maturity maintenance plus development (or reproduction).

The general philosophy behind the DEB theory is that there is a full balance approach for food (nutrients, energy, etc): "what goes in shall come out". Offspring is (indirectly) produced from food, which relates reproduction to feeding. Large individuals eat more than small ones, which links feeding to growth. Maintenance represents a drain of resources that is not linked to net synthesis of tissue or to reproduction. An increase of maintenance, therefore, indirectly leads to a reduction of growth, so to a reduction of feeding and reproduction.

This reasoning shows that the model requires a minimum level of complexity to address the various modes of action of a compound. One needs to identify this route to translate effects on individuals to that on the growth of natural populations (in the field). If food conditions are good, investment in maintenance, for instance, comprises only a small fraction of the daily food budget of individuals. Small effects of a toxicant on maintenance, therefore, result in very small effects on the population growth rate. If food conditions are poor, however, maintenance comprises a large fraction of the daily food budget. Small effects on maintenance can now translate into substantial effects on the population size. This reasoning shows that effects on populations depend on food conditions, which generally vary in time [Kooijman (1985, 1988), Hallam *et al.* 1989]. The different modes of action usually result in very similar point estimates for the NEC, within the current experience. Furthermore, no effects on individuals implies no effects on populations of individuals, but the mode of action is particularly important for predicting the effects at the population level.

7.2.4 Change in target parameter

The value of the target parameter is assumed to be linear in the internal concentration. The argumentation for this very simple relationship is in the Taylor's Theorem which states that any regular function can be approximated with any degree of accuracy for a limited domain by a polynomial of sufficiently large order. The interest is usually in small effects only, and routine applicability urges for maximum simplicity, so a first order polynomial (i.e. a linear relationship) is a strategic choice.

The biological mechanism of a linear relationship between the parameter value and internal concentration boils down to the independent action at the molecular level. Each molecule that exceeds an individual's capacity to repress effects acts independent of the other molecules. Think of the analogy where photosynthesis of a tree is just proportional to the number of leaves as long as this number is small; as soon as the number grows large, self-shading occurs and photosynthesis is likely to be less than predicted.

We doubtlessly require nonlinear responses for larger effect levels, but then also need to include more types of effects. Interesting extensions include receptor-mediated effects. The biochemistry of receptors is rather complex.

Two popular models are frequently used to model receptor-mediated effects and concentration:

a) the Michaelis Menten model boils down to a hyperbolic relationship, rather than a linear one which has one parameter more (Muller and Nisbet 1997);

b) the Hill model boils down to a log-logistic relationship [and has two parameters more than the linear model (Hill 1910, Garric *et al.* 1990, Vindimian *et al*. 1983).

Such extensions are particularly interesting if toxico-kinetics is fast, and the internal concentration is proportional to the external one (such as in cell cultures). The assumption that the target parameter is linear in the internal concentration does *not* translate into a linear response of the endpoint; it usually translates into sigmoid concentration-endpoint relationships, which are well known from empirical results. Notice that the linear model is a special case of the hyperbolic one, which is a special case of the log-logistic one.

7.2.5 Change in endpoint

The DEB model specifies how changes in one or more target parameters translate into changes in a specified endpoint. Popular choices for endpoints are reproduction rates (number of offspring per time), cumulative number of offspring (in daphnia-reproduction tests), body length (in fish-growth tests) and survival probability. Survival and reproduction together determine steady-state population growth, if they are known for all ages. Reproduction rates depend on age, namely, and the first few offspring contribute much more to population growth than later offspring. This is a consequence of the principle of interest-upon-interest; early offspring start reproduction earlier than later offspring. As is discussed below, indirect effects on reproduction come with a delay of the onset of reproduction, while direct effects on reproduction do not. The DEB model takes care of these more complex, but important, aspects of reproduction. Given the DEB model, there is no need to study all ages of the test organism once the DEB parameters are known. This application requires some basic eco-physiological knowledge about the species of test organism, but the acquisition of this knowledge does not have to be repeated for each toxicity test.

7.3 Survival

7.3.1 Relationship between hazard rate and survival probability

The effects on the survival probability of individuals are specified via the hazard rate. A hazard rate (dimension: probability per time) is also known as the instantaneous death rate. The hazard rate, $h(t)$, relates to the survival probability, $P_{\text{surv}}(t)$, as

$$h(t) = -P_{\text{surv}}(t)^{-1} \frac{\mathrm{d}}{\mathrm{d}t} P_{\text{surv}}(t)$$

or

$$P_{\text{surv}}(t) = \exp\left[-\int_0^t h(s) \ \mathrm{d}s \right] .$$

The product h times dt has the interpretation of the probability of dying in a small time increment, dt, given that the organism is alive at time t.

If the hazard rate is constant, which is the standard assumption for the death rate in the control, the relationship between the survival probability and the hazard rate reduces to

$$P_{\text{surv}}(t) = \exp[-h(t)].$$

Generally, the hazard rate increases with time, however. The mortality process can be modelled via the hazard rate, as is standard in survival analysis (Miller 1981; Cox and Oakes 1984). The hazard rate can depend on ageing and toxicity, as implied by the present model for survival, and can decrease in time, if, for instance, the concentration of a toxic compound decreases in time. If the concentration is constant, the ultimate LC_{50} equals the NEC.

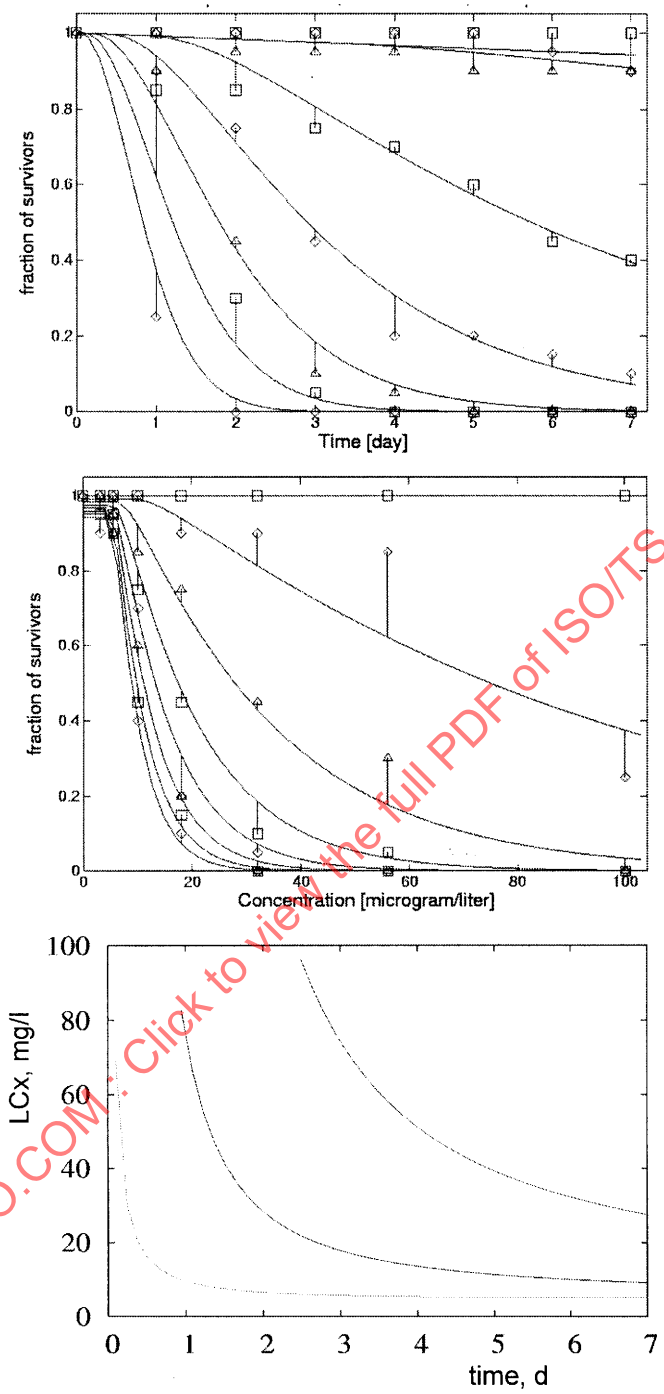
7.3.2 Assumptions of survival probability at any concentration of test compound

The following assumptions specify the survival probability at any concentration of test compound:

- assumptions on control behaviour
  - the hazard rate in the control is constant;
  - the organisms do not grow during exposure;
- assumption on toxico-kinetics
  - the test chemical follows first order kinetics;
- assumption on effects
  - the hazard rate is linear in the internal concentration;
- assumptions on measurements/toxicity test
  - the concentrations of test-compound are constant during exposure;
  - the measured numbers of dead individuals in subsequent time intervals are independently multinomially distributed.

7.3.3 Summary

In summary, the model amounts to: the hazard rate is linear in the internal concentration, which follows first order kinetics. These assumptions result in sigmoidal concentration-survival relationships, not unlike the log-logistic one, with a slope that increases during exposure (see Figure 22).



NOTE The resulting ML estimates are control hazard rate = 0,008 3 1/d NEC = 5,2 µg/l; killing rate 0,037 $(\mu g.d)^{-1}$, elimination rate = 0,79 d^{-1}. From the last three parameters, LC_x-time curves can be calculated; curves for the LC_0, LC_{50} and LC_{99} are shown (calculated with DEBtox and DEBtool, see 7.9). For long exposure times, the LC_x curves tend towards the NEC, for all x, in absence of blank mortality.

Figure 22 — Time and concentration profiles of the hazard model, together with the data of Figure 27

As is shown, the three exposure-time-independent parameters of the hazard model completely determine the response surface, thus the LC_x-time curves. It is even possible to reverse the reasoning.

EXAMPLE 1

If $LC_{50(1d)} = 50$ mM, $LC_{50(2d)} = 30$ mM and $LC_{50(3d)} = 25$ mM,

then NEC = 17,75 mM, the killing rate = 0,045 1/(mM.d) and the elimination rate = 2,47 1/d.

Such reconstructions are not very reliable, however, but they improve somewhat if more LC_{50} values are used.

If the observation times are very close together, the resulting huge matrix of survival-count data can be reduced to time-to-death data. Concentration-response modelling is traditionally considered to be different from time-to-death modelling, c.f. Newman *et al.* (1989), Dixon and Newman (1991), Diamond *et al.* (1991), but in the framework of biology-based models, these two approaches are just extreme cases of analyses of response-surfaces; their distinction vanishes and we generally deal with mixtures of both. The log-likelihood function then reduces to

$$l = \sum_i \ln\ h\ (t_i) - \sum_j \int_0^{t_j} h(s)\mathrm{d}s$$

where the first summation is across the individuals that actually died at the observed time points, excluding the ones that are taken alive out of the experiment. This can happen, for instance at the end of the experiment, or because their internal concentration is measured in a destructive way.

The second summation is across all individuals (the ones that died, as well as the ones that were removed alive). This sampling scheme allows that the concentrations for all individuals differ.

EXAMPLE 2 An example of this application is as follows:

Time-to-death and concentration pairs (in d and mM, respectively):

(21;1); (20;1,1); (20;0,9); (18;1,2); (16;1,3); (16;1,4); (15;1,5); (10;2); (9;1,8); (6;2,2); (5;2,5); (2;3); (2;4,3); (1;5); (1;4,5).

Time-of-removal and concentration pairs:

(21;0); (21;0); (21;0); (21;1).

The ML estimates for this combined data set for 19 individuals in total are

control hazard rate = 0,061 d^{-1},

NEC = 1,93 mM,

killing rate = 0,33 1/(mM.d),

elimination rate 0,75 d^{-1}.

This means, for instance, that the $LC_{50(2d)}$ = 5,6 mM and the $LC_{50(21d)}$ = 2,06 mM. (Calculations were made with DEBtool, see 7.9.3.)

The link between the DEB theory and the survival model is in the aging module of the DEB model, where the hazard rate, as affected by the ageing process, depends on the respiration rate in a particular way due to the action of free radicals; genotoxic compounds have a very similar mode of action and these compounds accelerate the ageing process (Kooijman 2000). The processes of tumour induction and growth have direct links with the ageing process (van Leeuwen and Zonneveld 2001). These effects on survival are beyond the scope of the present document, which deals with survival during (short) standardized exposure experiments.

On the assumption that test animals do not recover from immobilization, the concept “death” can be replaced by “initiation of immobilization” in this model. Due to the nonlinearity that is inherent to toxico-kinetics, this model does not belong to the class of generalized linear models for survival, which has been proposed for the analysis of toxicity data (Newman 1995, McCullagh and Nelder 1989).

The model for effects on survival, and details about the statistical properties of parameter estimates (especially that of NECs) are discussed in Andersen *et al*. (2000), Bedaux and Kooijman (1994), Klepper and Bedaux (1997, 1997a), Kooijman and Bedaux (1996, 1996a). Effects at time-varying concentrations are discussed in Péry *et al*. (2001, 2002), Widianarko and van Straalen (1996).

7.4 Body growth

7.4.1 Routes for affecting body growth

The DEB model allows for (at least) three routes for affecting body growth:

a) **a decrease of the assimilation rate**: Assimilation deals with the transformation from food into reserves, and can be affected by a decrease of the feeding rate, or a decrease of the digestion efficiency;

b) **an increase of the somatic maintenance costs**: These costs comprise protein turnover, the maintenance of intracellular and intra-organismal concentration gradients of compounds, osmo-regulation, heating of the body (mainly in birds and mammals), activity, and other drains on resources that are not linked to processes of net synthesis. Somatic maintenance costs directly compete with body growth for resources (in the DEB model). Thus an increase of maintenance costs directly results in a decrease of body growth, due to conservation of mass and energy;

c) **an increase in the specific costs for growth**: This is the case where the resource allocation to body growth is not affected, but the conversion of these resources to new tissue is.

This list does not exhaust all possibilities. An interesting alternative is in the change of the allocation to somatic maintenance plus body growth versus maturity maintenance and maturation (or reproduction). Under control conditions, the DEB model takes the relative investments in these two destinations to be constant (the absolute investments can change in time). Parasites and endocrine disrupting compounds (e.g. Andersen *et al.* 2001, Kooijman 2000) are found to change these relative investments. It is possible that a large number of compounds have similar effects. A practical problem in the application of a model that accounts for changes in the allocation fraction is that standardized tests for body growth do not include measurements that are necessary to quantify the effect appropriately. Detailed modelling of effects on mammalian development has been developed and applied (Setzer *et al*. 2001, Lau *et al.* 2000), but such approaches require adequate data and are specific for the compound as well as the test organism.

7.4.2 Assumptions

The following assumptions specify the effect on body growth at any concentration of test compound:

- assumption on control behaviour
  - the test-organisms follow a von Bertalanffy growth curve in the control;
- assumption on toxico-kinetics
  - the test chemical follows first order kinetics
    (Dilution by growth is taken into account);
- assumption on effects

  One of three modes of action occur:
  - the assimilation rate decreases linearly in the internal concentration;
  - the maintenance rate increases linearly in the internal concentration;
  - the costs for growth increase linearly in the internal concentration;
- assumptions on measurements/toxicity test
  - the concentrations of test-compound are constant during exposure;
  - the measured body lengths are independently normally distributed with a constant variance.

7.4.3 Von Bertalanffy growth curve

The von Bertalanffy growth curve is given by

$$L(t) = L_\infty - (L_\infty - L_0)\exp\{-r_b t\},$$

where

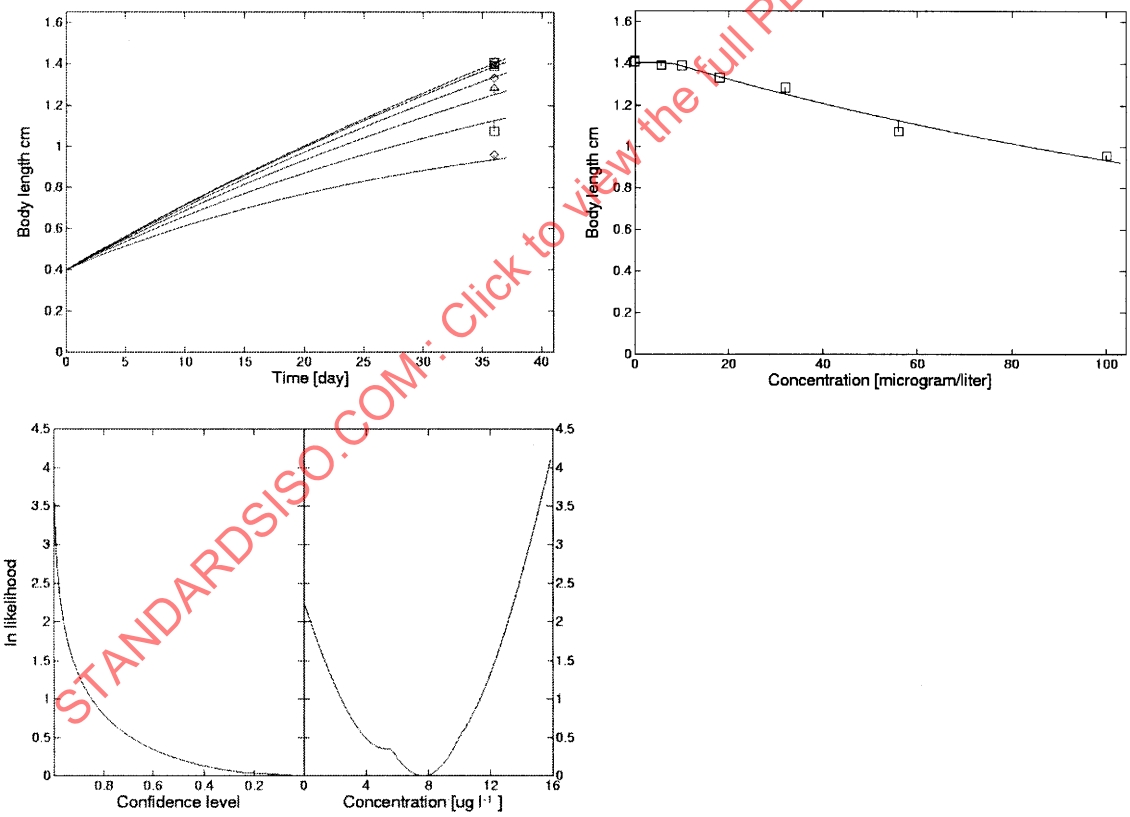
$L(t)$ is the length at time, t;

L_0 is the initial length;

L_∞ is the ultimate length; and

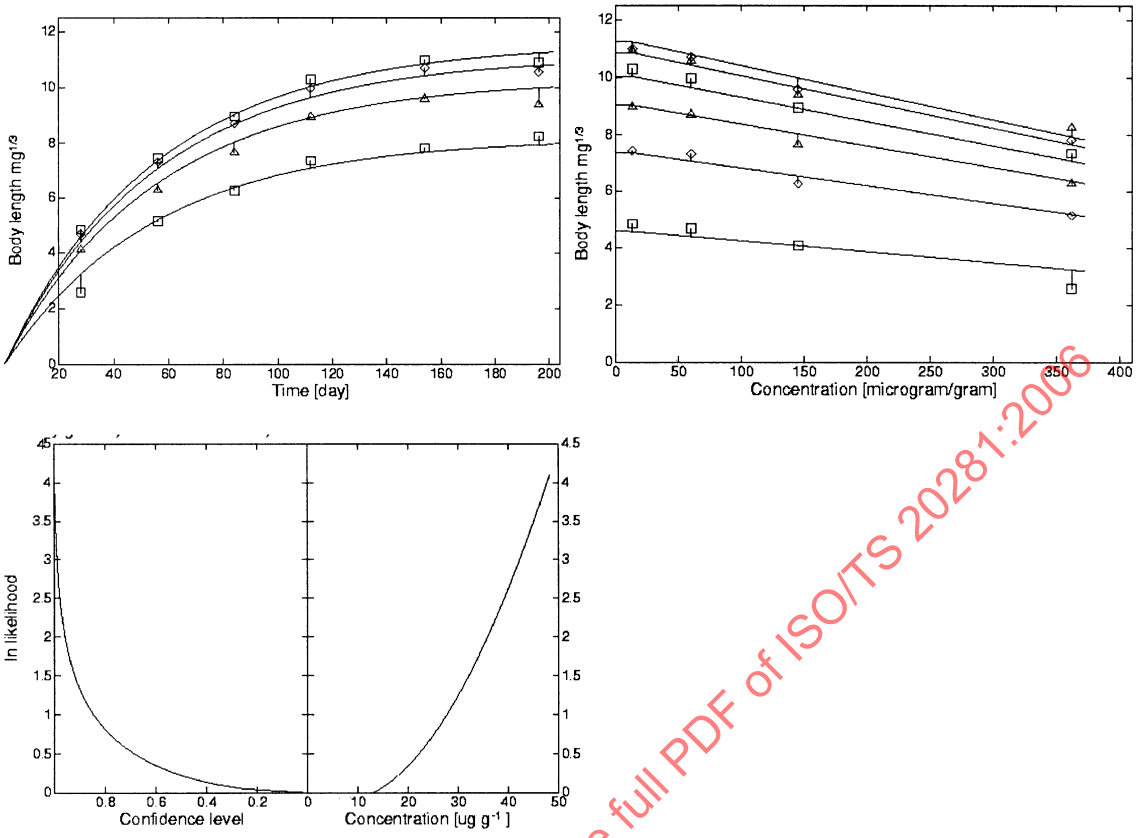
r_b is the von Bertalanffy growth rate.

The DEB model predicts that body growth is of the von Bertalanffy type only at constant food densities, in the case of isomorphs (i.e. organisms that hardly change in shape during growth). An implied assumption is, therefore, that food density is constant, or high. Food intake depends hyperbolically on food density in the DEB model; variations in food density, therefore, hardly result in variations in food intake as long as food remains abundant. Examples of application of the model of effects on growth by an increase of the maintenance costs and by a decrease of assimilation are shown in Figures 23 and 24.



NOTE The parameters estimates are NEC = 7,65 g/l; control ultimate length = 37 mm; tolerance concentration = 43,5 g/l; elimination rate = large. Fixed parameters are initial length = 4 mm; von Bertalanffy growth rate = 0,01 d. The profile likelihood function for the NEC is given left. The $EC_{0(36d)}$ = 766 g/l; $EC_{50(36d)}$ = 176 g/l. The use of the profile likelihood graphs to obtain confidence intervals is explained in the legend to Figure 29.

Figure 23 — Time and concentration profiles for effects on growth of *Pimephalus promelas* via an increase of specific maintenance costs by sodium pentachlorophenate (data by Ria Hooftman, TNO-Delft)



NOTE

The parameter estimates are NEC = 13 g/g; control ultimate length = 11,6 mm; tolerance concentration = 1,2 mg/g; elimination rate = large.

Fixed parameters are initial length = 0 mm; von Bertalanffy growth rate = 0,018 d. The profile likelihood function for the NEC is given in the bottom left graph. $EC_{0(100d)}$ = 13 g/g; $EC_{50(100d)}$ = 605 g/g.

Figure 24 — Time and concentration profiles for effects on growth of *Lumbricus rubellus* via a decrease of assimilation by copper chloride (data from Klok and de Roos 1996)

The first example (Figure 23) shows that it is not necessary to have observations in time; the second example (Figure 24) shows that it is not absolutely necessary to have a control. Although inclusion of a control is always advisable, the control is treated in the same way as positive concentrations in the DEBtox method. The statistical properties of the parameter estimates and the confidence one has in them obviously improve if controls and positive concentrations are available.

At high concentrations, the test compound probably not only affects body growth, but usually also survival. The DEBtox software (see 7.9) accounts for differences in number of individuals of which the body size has been measured.

The models for effects on body growth, and details about the statistical properties of parameter estimation (especially that of NECs, are discussed in Kooijman and Bedaux (1996, 1996a).

7.5 Reproduction

7.5.1 Routes that affect reproduction

The DEB model allows for (at least) five routes that affect reproduction. The first three routes are identical to that for growth and are called the indirect routes. The DEB model assumes namely that food intake is proportional to surface area, so big individuals eat more than small ones. This means that if growth is affected,

feeding is directly or indirectly affected as well, which leads to a change in resources that are available for reproduction. The routes not only lead to a reduction of reproduction, but also to a delay of reproduction. In addition, there are two direct routes for affecting reproduction:

a) an increase in the costs per offspring, so an effect on the transformation from reserves of the mother to that of the embryo;

b) death of early embryos, before they leave the mother. Dead embryos can be born, or are absorbed; only the living ones are counted.

These two direct routes assume that the allocation to reproduction is not affected by the compound, but that the compound affects the conversion of these resources into living embryos.

7.5.2 Assumptions

The following assumptions specify the effect on reproduction at any concentration of test compound:

- assumptions on control behaviour
  - the test-organisms follow a von Bertalanffy growth curve in the control;
  - reproduction depends on assimilation, maintenance and growth as specified by the Dynamic Energy Budget (DEB) theory;
- assumption on toxico-kinetics
  - the test chemical follows first-order kinetics
    (Dilution by growth is taken into account.);
- assumptions on effects

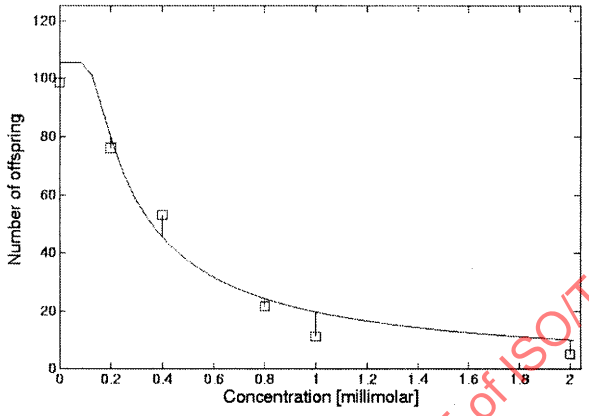
  One of five modes of action occur:
  - the assimilation rate decreases linearly in the internal concentration;
  - the maintenance rate increases linearly in the internal concentration;
  - the costs for growth increase linearly in the internal concentration;
  - the costs for reproduction increase linearly in the internal concentration;
  - the hazard rate of the neonates increases linearly in the internal concentration;
- assumptions on measurements/toxicity test
  - the concentrations of test-compound are constant during exposure;
  - the measured cumulative numbers of young per female are independently normally distributed with a constant variance.

7.5.3 Implication

An implication of the DEB theory is that indirect effects on reproduction (the first three modes of action) are a reduction of the reproduction rate as well as a delay of the start of reproduction, while direct effects (the last two modes of action) involve a reduction of reproduction only. All three indirect effects on reproduction also have effects on growth, despite the fact that just a single target parameter is affected. The delay of the onset of reproduction is, therefore, coupled to effects on growth. The measurement of body lengths at the end of the test on reproduction can be used as an easy check and as an identification aid to the mode of action. This

mode of action is of importance to translate effects on individuals into those on growing populations (Kooijman 1985, Nisbet *et al.* 2000).

The DEBtox software (see 7.9) accounts for possible reductions of numbers of survivors in the reproduction test via weight coefficients; the more females contribute to the mean reproduction rate per female, the more weight that data point has in the parameter estimation. An example of application is from the OECD ring-test for effects of cadmium on Daphnia reproduction (Figure 25); the full results are reported in Kooijman *et al.* (1998):



NOTE The figures show the time and concentration profiles.

The parameter estimates are NEC = 3,85 nM; tolerance concentration = 5,40 nM; max reproduction rate = 14,4 d; elimination rate = 3,0 d.

Fixed parameters are von Bertalanffy growth rate = 0,1 1/d; scaled length at birth = 0,13; scaled length at puberty = 0,42; energy investment ratio = 1. The NEC does not differ significantly from 0 on the basis of these data. If a more accurate estimate is required, lower test concentrations should be selected. These parameter values imply $EC_{0(21d)}$ = 0,1 mM and $EC_{50(21d)}$ = 0,336 mM.

Figure 25 — Effects of cadmium on the reproduction of Daphnia magna through an increase of the costs per offspring — Data from the OECD ring-test

The models for effects on reproduction, and details about the statistical properties of parameter estimation (especially that of NECs) are discussed in Kooijman and Bedaux (1996b, 1996c).

7.6 Population growth

7.6.1 General

If individuals follow a cycle of embryo, juvenile and adult stages, one needs the context of physiologically structured population dynamics to link the behaviour of population dynamics to that of individuals. If the individuals only grow and divide, a substantial simplification is possible in the context of the DEB model. This is the case in the algal growth inhibition tests, and in tests with duckweed, for instance.

7.6.2 Assumptions

Three modes of action of the compound are delineated here. The following assumptions specify the model for effects on populations:

— assumptions on control behaviour

  — the viable part of the population grows exponentially (the cultures are not nutrient or light limited during the test);

— assumption on toxico-kinetics

  — the internal concentration is rapidly in equilibrium with the medium;

— assumptions on effects

  One of three modes of action occur:

  — the costs for growth are linear in the (internal) concentration;

  — the hazard rate is linear in the (internal) concentration during a short period at the start of the experiment;

  — the hazard rate is linear in the (internal) concentration during the experiment;

— assumptions on measurements/toxicity test

  — the concentrations of test-compound are constant during exposure;

  — the inoculum size is the same for all experimentally tested concentrations;

  — biomass measurements include living and dead organisms;

  — the measured population sizes are independently normally distributed with a constant variance.

The rationale of the second mode of action (death only at the start of the experiment) is that effects relate to

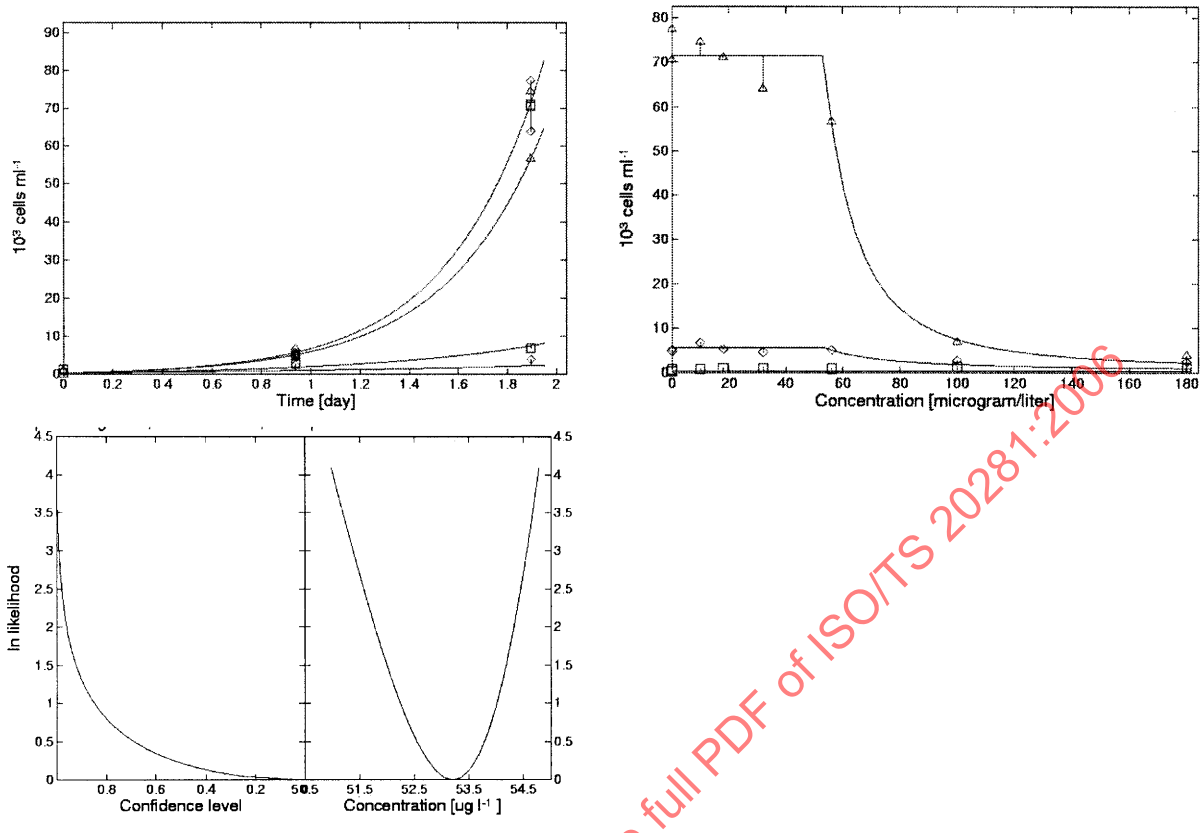
— the transition from control culture to stressed conditions, not to the stress itself;

— the position of the transition in the cell cycle. Cells are not synchronized, so the transition occurs at different moments in the cell cycle, for the different cells. If cells are more sensitive for the transition during a particular phase in the cell cycle, only those cells are affected that happen to be in that phase.

The EC_x values for this type of test can be calculated in various ways, with different results. One way to do this is on the basis of biomass as a function of time. This should not be encouraged, however because the result depends on experimental design parameters that have nothing to do with toxicity (Nyholm 1985). Another way to do this is on the basis of specific population growth rates, which are independent of time (Kooijman *et al.* 1996a). An example of application of the DEBtox method is given in Figure 26.

	Time: day, Conc: microgram/liter, Resp: 10^3 cells ml^{-1}							
	0	0	10	18	32	56	100	180
0.0000	0.4	0.8	0.9	1.1	1.1	1.0	1.4	1.1
0.9375	4.9	5.4	6.8	5.4	4.7	5.2	2.8	2.5
1.8958	70.5	77.4	74.5	71.1	64.0	56.6	6.9	3.9

Figure 26 — Example of application of the DEBtox method



NOTE The figures show the data and the time and concentration profiles (note that this data set contains two blanks). The estimated parameters are inoculum = 494 cells/ml; specific growth rate = 2,62 1/d; NEC = 0,053 mg/l; tolerance concentration = 0,0567 mg/l.

The profile likelihood function for the NEC is given in the lower left-hand figure. The EC_{50} = 0,062 4 mg/l. The robustness of this approach is demonstrated by the fact that removal of the highest concentration leads to the same point estimate for the NEC (but with a larger confidence interval).

Figure 27 — The effect of a mixture of C,N,S-compounds on the growth of *Skeletonema costatum* via an increase of the costs for growth — Data from the OECD ring test

The model for effects on population growth, and details about the statistical properties of parameter estimation (especially that of NECs) are discussed in Kooijman *et al*. (1996a). Toxic effects on logistically growing populations in batch cultures are discussed in Kooijman *et al*. (1983); a paper on the interference of toxic effects and nutrient limitation is in preparation. (Also see Figure 27.)

7.7 Parameters of effect models

7.7.1 General

The parameters of effect models can be grouped into a set that relates directly to the effects of the test compound and a set that relates to the eco-physiological behaviour of the test organisms.

7.7.2 Effect parameters

7.7.2.1 Toxicity and dynamic parameters

The basic biology-based models have two toxicity parameters and a single dynamic parameter:

— **NEC** = $EC_0(\infty)$: No effect concentration, which is the 0 % effect level at very long exposure times (dimension: external concentration);

— **killing rate** (for effects on survival; dimension: per external concentration per time) or tolerance concentration (for sub-lethal effects; dimension: external concentration);

— **elimination rate** of first-order kinetics (for survival, body growth and reproduction tests; not for population growth inhibition tests. Dimension: per time). Large values mean that the internal concentration rapidly reaches equilibrium with the concentration in the medium. If the internal concentration is in equilibrium, the effects no longer change. Notice that the elimination rate has no information about the toxicity of the test compound.

7.7.2.2 Killing rate, b_k

The **killing rate** is the increase in the hazard rate per unit of concentration of test compound that exceeds the NEC:

$$h(t) = h_{\text{control}} + b_k \left(\frac{c_{\text{int}}}{k_{\text{BCF}}} - c_0 \right)_+$$

where

$h(t)$ is the hazard rate, at time, t;

h_{control} is the control hazard rate;

b_k is the killing rate;

c_0 is the no effect concentration;

c_{int} is the internal concentration;

k_{BCF} is the bio-concentration factor; and

$+$ means that, if $\frac{c_{\text{int}}}{k_{\text{BCF}}}$ is below c_0, then the hazard rate equals the control hazard rate.

The k_{BCF} stands for the ratio of the internal and external concentrations *in equilibrium*. No assumptions are made about its value; it can be very small for compounds that hardly penetrate the body.

The **tolerance concentration** quantifies the change in the target parameter per unit of concentration of test compound that exceeds the NEC:

The parameter value is equal to the control parameter value times (1 plus the stress value).

$$s(c_q) = \frac{1}{c_*} \left(\frac{c_{\text{int}}}{k_{\text{BCF}}} - c_0 \right)_+$$

where

$s(c_q)$ is the stress function;

c_* is the tolerance concentration;

c_{int} is the internal concentration; and

k_{BCF} is the bio-concentration factor.

The target parameter value in this specification of the tolerance concentration can be the specific costs for growth, the specific maintenance costs or another physiological target parameter. This depends on the mode of action of the compound.

The name “tolerance concentration” refers to the fact that the higher its value, the less toxic the chemical compound. Notice that the ratio “$\frac{c_{\text{int}}}{k_{\text{BCF}}}$” has the interpretation of an external concentration that is proportional to the internal concentration; the tolerance concentration, like the NEC, has the dimension of an external concentration. This is done because internal concentrations are generally unknown in practice. The internal concentration, and so the stress value, depends on the (constant) external concentration and the (changing) exposure time. The stress value is a dimensionless quantity, which is only introduced to simplify the specification of the change in the target parameter.

7.7.3 Discussion

The NEC, the elimination rate and the tolerance concentration (or killing rate) are parameters that do NOT depend on the exposure time. This is in contrast to EC_x values, which do depend on exposure time. Notice that the accumulation rate (a toxico-kinetic parameter) does not occur in the parameter set of effect models. This is because less toxic compounds that accumulate strongly cannot be distinguished from toxic compounds that hardly accumulate if only effects, and no internal concentrations are observed. This is also the reason why NECs, killing rates and tolerance concentrations are in terms of external concentrations, while the mechanism is via internal concentrations. Effect models treat internal concentrations as hidden variables.

The kinetic parameters depend on the properties of the chemical compound. The elimination rate is inversely proportional to the square-root of the octanol-water partition coefficient (P_{ow}), while the uptake rate is proportional to the square-root of this coefficient (Kooijman and Bedaux 1996, Kooijman 2000). Since effects depend on internal concentrations, so on toxico-kinetics, effect parameters depend on the partition coefficient as well; the NEC, tolerance concentration and inverse killing rate are all inversely proportional to the P_{ow} (Gerristen 1997, Kooijman and Bedaux 1996, Kooijman 2000). Such relationships can be used in practice to test parameter estimates against expectations.

The prediction of how the toxicity parameters depend on the octanol-water partition coefficient can be used for selecting appropriate concentrations to be tested.

EXAMPLE 1 An example is as follows.

Suppose that compound 1 with $P_{\text{ow}} = 10^6$ has been tested for its effects on survival, which resulted in the parameter estimates NEC = 1,3 mM; killing rate = 1,5 1/(mM.d); elimination rate = 0,5 1/d.

Now we have to test compound 2, with a physiologically similar mode of action and a $P_{\text{ow}} = 10^7$.

We expect to find the parameter estimates

NEC = 0,13 mM; killing rate = 15 1/(mM.d); elimination rate = 0,5/√10 = 0,16 1/d.

These three parameters imply that the $LC_{0(2d)}$ = 0,47 mM and the $LC_{99(2d)}$ = 1,9 mM, which gives some guidance for choosing the concentration range to be tested in a test of 2 d.

Suppose now that we tested compound 1 for effects on reproduction in Daphnia with a control max reproduction rate of 15 offspring per day.

Let us assume that the compound increases the maintenance costs.

This resulted in NEC = 1,3 mM; tolerance concentration = 10 mM; elimination rate = 0,5 1/d.

We expect to find for compound 2

NEC = 0,13 mM; tolerance concentration = 1 mM; elimination rate = 0,16 1/d.

These three parameters imply that the $EC_{0(21d)}$ = 0,18 mM and the $EC_{99(21d)}$ = 1,9 mM, which gives some guidance for choosing the concentration range to be tested in a reproduction test of 21 d.

(Calculations with DEBtool, see 7.9.3.)

Contrary to the more usual techniques to establish Quantitative Structure Activity Relationships (QSARs), the influence of the P_{ow} on the parameters of biology-based models can be predicted on the basis of first principles; these QSARs are not derived from regression techniques that require toxicity data for other compounds. The reason why traditional regression techniques for establishing QSARs are somewhat cumbersome is in the standardization of the exposure period. For any fixed exposure period (usually 2 d or 14 d), the LC_{50} (or EC_{50}) for a compound with a low P_{ow}, P_{ow} is close to its LC_{50} for very long exposure times. For compounds with a large P_{ow}, however, the ultimate LC_{50} is much lower than the observed one. If we compare LC_{50}s for low P_{ow} and high P_{ow} values, we observe complex deviations from simple relationships, which are masked in log-log plots and buried in the allometric models that are usually applied to such data. (An allometric model is a model of the type $y(x) = a\,x^b$, where a and b are parameters.)

Effects of modifying factors, such as pH, can be predicted, and taken into account in the analysis of toxicity data (corrections on measured or nominal concentrations, and on measured or modelled pH values). If the compound affects the pH at concentrations where small effects occur, and the NEC and/or the killing rate of the molecular and ionic forms differ, the following relationships apply:

$$b_k(pH) = \frac{b_k^m + b_k^i 10^{pH-pK}}{1+10^{pH-pK}} \quad \text{and} \quad c_0(pH) = c_0^i c_0^m \frac{1+10^{pH-pK}}{c_0^i + c_0^m 10^{pH-pK}}$$

where

pK is the ion-product constant;

c_0^i is the NEC of the ionic form;

c_0^m is the NEC of the molecular form;

b_k^i is the killing rate of the ionic form; and

b_k^m is the killing rate of the molecular ionic form (Kooijman 2000, Könemann 1980).

The pH is affected much more easily in soft than in hard water (see e.g. Segel 1976, Stumm and Morgan 1996). Compounds may affect internal pH to some extent; in that case the relationship is approximately only.

On the assumption that the chemical environment inside the body of the test organisms is not affected (due to homeostatic control), the observed survival pattern can be used to infer about the toxicity of the molecular and the ionic form. The partitioning between the molecular and ionic form is fast, relative to the uptake and elimination (both in the environment and in the organism); this means that the elimination rate relates to both the molecular and the ionic form. An example is given in Table 7.

Table 7 — Example of a pH effect on mortality

pH	7,5	7,5	7,4	7,2	6,9	6,6	6,3	6,0
Concentration	0	3,2	5,6	10	18	32	56	100
Days	**Number of survivors**							
0	20	20	20	20	20	20	20	20
1	20	20	20	20	20	20	19	18
2	20	20	19	19	19	18	18	18
3	20	20	17	15	14	12	9	8
4	20	18	15	9	4	4	3	2
5	20	18	9	2	1	0	0	0
6	20	17	6	1	0	0	0	0
7	20	5	0	0	0	0	0	0

Suppose that we found the numbers of survivors as in the left table for a compound with ionisation product constant of 9,0. The parameter estimates are (calculations with DEBtool, see 7.9.3) given in Table 8.

Table 8 — Example of estimated DEBtool parameters

	Molecule		Ion	
	ML [a]	SD [b]	ML [a]	SD [b]
Control mortality rate	0,009	0,005		
NEC	24,9	16,9	0,17	0,03
Killing rate	0,039	0,013	2,82	2,16
Elimination rate	1,48	0,50		

[a] ML is the maximum likelihood.

[b] SD is the standard deviation.

The elimination rate is proportional to the ratio of a surface area and a volume of the test organism, which yields an inverse length measure. This relationship implies predictable differences between elimination rates in organisms of different sizes, which have been tested against experimental data (see e.g. Gerritsen 1997). This is rather straightforward in the case of individuals of the same species, but also applies to individuals of different, but physiologically related, species. The body size scaling relationships as implied by the DEB theory suggest predictable differences in the chemical body composition, in lipid content and in elimination rate and toxicity parameters. Such relationships still wait for testing against experimental data, but are helpful in developing an expectation for parameter values; such expectations can be used in experimental design, and in checking results of parameter estimations.

The prediction of how the three parameters of the hazard model depend on the body size of the test organisms can also be used for selecting appropriate concentrations to be tested.

EXAMPLE 2

Suppose that a compound has been tested using fish of a mass of 1 mg, which resulted in the parameter estimates

NEC = 1,3 mM; killing rate = 1,5 1/(mM.d); elimination rate = 0,5 1/d.

Now we have to test the compound for fish of 1 g of the same species.

We expect to find a difference in the elimination rate only, i.e. 0,5/10 = 0,05 1/d. These three parameters imply that the $LC_{0(2d)}$ = 1,4 mM and the $LC_{99(2d)}$ = 5,5 mM, which gives some guidance for choosing the concentration range to be tested in a test of 2 d.

(Calculations were made with DEBtool, see 7.9.3.)

7.7.4 Eco-physiological parameters

The model for effects on survival has the **control mortality rate** as a parameter, which results in an exponentially decaying survival probability. This means that the model delineates two causes for death: death due to background causes (for instance manipulation during the assay) and death due to the compound. This obviously complicates the analysis of the death rate at low exposure levels, because we can never be sure about the actual cause of death in any particular case. Not only the data in the control, but all data are used to estimate the control mortality rate; if no death occurs in the control, this does not imply that the control mortality rate is zero. The profile likelihood function for the NEC quantifies the likelihoods of the two different causes of death. Figures 22, 23 and 24 show how background causes can be distinguished from those by the compound.

	Time: day, Conc.: microgram/liter							
	0.0	3.2	5.6	10.0	18.0	32.0	56.0	100.0
0	20	20	20	20	20	20	20	20
1	20	20	20	20	18	18	17	5
2	20	20	19	17	15	9	6	0
3	20	20	19	15	9	2	1	0
4	20	20	19	14	4	1	0	0
5	20	20	18	12	4	0	0	0
6	20	19	18	9	3	0	0	0
7	20	18	18	8	2	0	0	0

NOTE The data in the body represent the number of surviving guppies. The first column specifies the observation times in days, the first row specifies the concentrations of dieldrin in g/l. Figure 29 shows how an answer can be found to the question whether the two deaths in the concentrations 3,2 g/l and 5,6 g/l are due to dieldrin, or to "natural" causes.

Figure 28 — A typical table of data that serves as input for the survival model, as can be used in the software package DEBtox (Kooijman and Bedaux 1996)

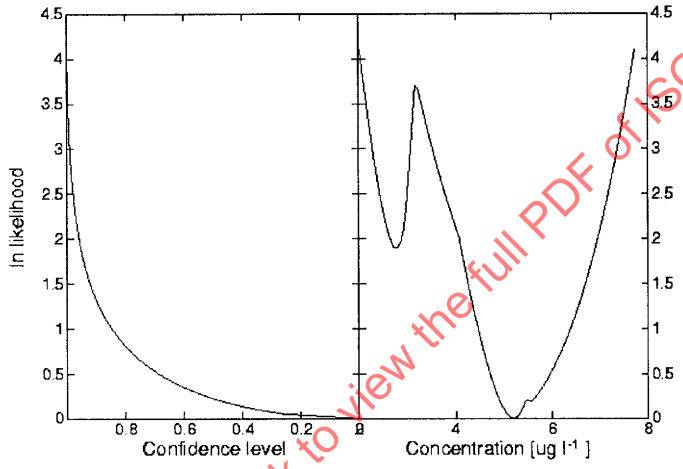


Figure 29 — This profile likelihood function of the NEC (right panel) for the data in Figure 28 results from the software package DEBtox (Kooijman and Bedaux 1996)

This profile determines the confidence set for the NEC (first select the confidence level of your choice in the left panel, then read the ln likelihood; the concentrations in the right panel for which the ln likelihoods are below this level comprise the confidence set of the NEC; the confidence set for the NEC is a single interval for low confidence levels, but a set of two intervals for high confidence levels). The maximum likelihood estimate for the NEC is here 5,2 g/l and corresponds to the interpretation of death in concentration 3,2 g/l due to "natural" causes; the second local extreme at 2,9 g/l corresponds to the interpretation of this death due to dieldrin. Figure 29 shows that this interpretation is less likely, but the figure shows that we cannot exclude this possibility for high confidence levels. If the lowest concentration would have no deaths in this data set, the profile likelihood function would not have a second local extreme.

The model for effects on growth have a single eco-physiological parameter each (the **ultimate body length**, and the **maximum reproduction rate**), that is estimated from the data, and a **scatter parameter** that stands for the standard deviation of the normally distributed deviations from the model predictions. The latter parameter also occurs in the models for effects on population growth.

The models for effects on body growth and reproduction have some parameter values that cannot be estimated from (routine) tests. Their values should be determined by preliminary eco-physiological experiments. These parameters are

— **von Bertalanffy growth rate** (dimension: per time). This parameter quantifies how fast the initial length approaches the ultimate length at constant food density. (The food density affects this parameter.) In principle, its value could be extracted from length measurements in the control, provided that enough observation times are included. Under standardized experimental conditions, its value should always be

the same, however. Moreover, the lengths are usually only measured at the end of the test only. These data do not have information about the value of the von Bertalanffy growth rate;

— **initial body length** (dimension: length), which is the body length at the start of the test. It is assumed that this applies to all individuals in all concentrations. The DEB model for reproduction has a scaled length at birth as parameter, which is dimensionless. This scaled length is the ratio of the length at birth and the maximum length of an adult at abundant food. Since the daphnia reproduction test uses neonates, the initial body length equals the length at birth;

— **scaled length at puberty** (dimensionless). This is the body length at the start of reproduction in the control as a fraction of the maximum body length of an adult at abundant food. The DEB model takes this value to be a constant, independent of the food density. At low food density, it takes a relatively long time to reach this length. The start of reproduction, therefore, depends on food density. The model for effects on reproduction needs the length at puberty. That on body growth does not use this parameter;

— **energy investment ratio** (dimensionless). This parameter stands for the ratio between the specific energy costs for growth and the product of the maximum energy capacity of the reserves and the fraction of the catabolic energy flux that is allocated to somatic maintenance plus growth. The maximum (energy) capacity of the reserves is reached after prolonged exposure to abundant food. The catabolic flux is the flux that is mobilized from the reserves to fuel metabolism (i.e. allocation to somatic and maturity maintenance, growth, maturation or reproduction; the relative allocation to somatic maintenance plus growth is taken to be constant in the DEB model). The value of the parameter does not affect the results in a sensitive way. The logic behind the DEB theory requires its presence, however; the parameter plays a more prominent role at varying food densities.

The DEBtox software (see below) fixes these parameters at appropriate default values for the standardized tests on fish growth and daphnia reproduction. The user can change these values.

The models for population growth have two eco-physiological parameters that are estimated from the data:

— the **inoculum size** (dimension: mass or number per volume), which is taken to be equal in all concentrations

— the control **specific population growth rate** (dimension: per time)

7.8 Recommendations

7.8.1 Goodness of fit

As applies to all models that are fitted to data, one should always check for goodness of fit (as incorporated in DEBtox), inspect the confidence intervals of the NEC, and mistrust any conclusion from models that do not fit the data (see also 6.4). The routine presentation of graphs of model fits is strongly recommended. “True” models, however, do not always fit the data well, due to random errors. If deviations between data and model-fits are unacceptably large, it makes sense to make sure that the experimental results are reproducible. Problems with solubility of the test compound, pH effects, varying concentrations, varying conditions of test animals, interactions between test animals and other factors can easily invalidate model assumptions. It might be helpful to realize that one approach for solving this problem is in taking such factors into account in the model (and apply a more complex model), but another approach is to change the experimental protocol such that the problems are circumvented. The models are designed to describe small effects; if the lack of fit relates to large effects, it can be recommended to exclude the high concentration(s) from the data analysis.

Any model might fit data well for the wrong reasons; a good fit does not imply the “validity” of that model. This should motivate to explore all possible means for checking results from data analysis; an expectation for the value of parameters is a valuable tool.

The assumption of first-order kinetics is not always realistic in detail. A general recommendation is to consider more elaborate alternatives only if data on toxico-kinetics are available. Depending on the given observation times, the elimination rate is not always accurately determined by the data. In such cases, one might consider to fix this parameter at a value that is extracted from the literature, and/or derived from a related compound, after correction for differences in P_{ow} values.

7.8.2 Choice of modes of action

Experience teaches us that the mode of action usually has little effect on the NEC estimates. Models for several modes of action frequently fit well to the same experimental data set; if additional type of measurements would have been available (such as feeding rate and/or respiration rate), it is much easier to choose between modes of action. These modes of action are of importance to translate effects on individuals to those on population dynamics, and how food availability interferes with toxic effects. The DEB theory deals with this translation.

Measurements of feeding and respiration rates, and of body size (in reproduction tests) greatly help identifying the mode of action of the compound. The proper identification of the mode of action is less relevant for estimates of the NEC.

7.8.3 Experimental design

DEBtox has been designed to analyse the results from toxicity tests as formulated in OECD guidelines (numbers 201, 202, 203, 204, 211, 215, 218, 219) and ISO guidelines (ISO 6341, ISO 7346-3, ISO 8692, ISO 10229, ISO 10253, ISO 12890, ISO 14669). The experimental design described in these guidelines is suitable for the application of DEBtox. Confidence intervals for parameter estimates are greatly reduced if not only the responses at the end of the toxicity experiments are used, but also observations during the experiment. Ideally, one should be able to observe how fast effects build up during exposure in the data, till the effect levels satiate. Note that this does not require additional animals to be tested, only that they are followed for a longer period of time.

Large extrapolations of effects, especially in the direction of longer exposure times, are generally not recommended; this is because, ideally, the assumptions need to be checked for all new applications. It, therefore, makes sense to let the optimal choice for the exposure period depend on the compound that is tested, and the test organisms that are used. The higher the solubility in fat of the test compound (e.g. estimated from P_{ow}), and the larger the body size of the test organisms, the longer the exposure should last.

As stated in the Introduction, it is strongly recommended to include all available observations into the analysis; not only those at the end of the experiment, but also the observations that have been collected during the experiment (for instance when the media are refreshed). It is generally recommended that the number of observations during exposure, the concentrations of test compound and the number of used test animals are such that the model parameters can be estimated within the desired accuracy.

Experimental design should optimize the significance of the test; the significance of single-species tests is discussed in Reference [232]. From a data analysis point of view, it makes sense to extend the exposure period till no further effects show up. The length of the exposure period then relates to the physical-chemical properties of the compound.

7.8.4 Building a database for raw data

Since biology-based methods not only aim at a description, but also at an understanding of the processes that underlie effects, it is only realistic to assume that this understanding evolves over the years. In the future, it might be useful to reanalyse old data in the light of new insights. In anticipation of this, it is recommended to build a raw database.

7.9 Software support

7.9.1 General

The models that are used by biology-based methods are fully derived and discussed in all mathematical detail in the open literature; a summary of the specification is given in Annex F of this report. There is, therefore, no need to use any of the software that is mentioned in this subclause. On the other hand, fitting sets of differential equations to data (as required by the models for effects on body growth and reproduction), the calculation of profile likelihoods for NECs, and the more advanced methods of fitting several datasets

simultaneously, is beyond the capacity of most standard packages. Even if packages can do the job, the optimization of numerical procedures (such as solving initial value problems) can be somewhat laborious.

The computations for biology-based methods have been coded in two packages, DEBtox and DEBtool, which can be downloaded freely from the electronic DEB-laboratory at http://www.bio.vu.nl/thb/deb/. Both packages are updated at varying intervals; the user has to check the website for the latest version. These packages are used in (free) international internet-courses that are organized by the Department of Theoretical Biology at the Vrije Universiteit, Amsterdam.

An MS Excel macro able to estimate Hill parameters using non-linear regression is available under the GPL license on the site: http://eric.vindimian.9online.fr/.

7.9.2 DEBtox

DEBtox is a load-module for Windows and Unix that is meant for routine applications.

The user cannot define new models. The package has many options for parameter estimation, confidence intervals and profile likelihoods (for the NEC for instance), fixation of parameters at particular values (such as NEC = 0) while estimating the other parameters, calculation of statistics (such as $EC_{x}.t$ and $ET_{x}.c$ values and their confidence intervals), hypothesis testing about parameter values (such as NEC ≠ 0), graphical representations to check goodness of fit, residual analysis, etc. Example data-files are provided for each toxicity test.

DEBtox is a user-friendly package, and the numerical procedures are optimized for the various models (modes of action) that can be chosen. The elimination rate, for instance, is not always accurately determined by the data, especially if a single observation time is given. DEBtox always calculates three sets of parameter estimates, corresponding with the elimination rate being a free parameter, or zero, or infinitely large. Only the best result is shown. The initial values for the parameters that are to be estimated are selected automatically. In fact, many trials (some hundred) are performed, and only the best result is shown. The user does not have to bother about these computational "details". (The likelihood function can have many local maxima, depending on the model and on the observations. The result of the numerical procedure to find a local maximum depends on the initial value; we are only interested in the global maximum, however. This problem complicates nonlinear parameter estimation in practice; it is an extra reason to check the result graphically in all applications.)

The current version of DEBtox can handle a single endpoint only (i.e. a single table of observations of responses at the various combinations of concentration and exposure time). In the period 2002-2006, DEBtox will be extended to include multiple samples to allow the analysis of effects on survival and reproduction simultaneously, and to test hypotheses about differences of parameter values between samples.

7.9.3 DEBtool

DEBtool is source code (in Octave and Matlab®) for Windows and Unix that is meant for research applications. Octave is freely downloadable, Matlab[21)] is commercial. DEBtool is much more flexible than DEBtox, but requires more knowledge for proper use; it is less user-friendly than DEBtox. Initial values for parameter estimations are not automatic, for instance. DEBtool has many domains that deal with the various applications of DEB models in eco-physiology and biotechnology; the domain "tox" deals with applications in ecotoxicology. The package can handle multiple data sets; several numerical procedures can be selected to find parameter estimates. DEBtool allows researchers to estimate parameters if the variance is proportional to the squared mean, to calculate the NEC, killing rate and elimination rate from LC_{50} values for three exposure times, to estimate parameters from time-to-death data, and to extract the toxicity parameters for the molecular and the ionic form when the pH is measured for each concentration, etc. Many specific models are coded, and the user can change and add models.

21) Matlab is an example of a suitable product available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of this products.

8 List of existing guidelines with references to the subclauses of this Technical Specification

Table 9 — Existing guidelines

Guideline/standard		Test	Endpoint	Reference (NOEC)	Reference (dose-response modelling)	Reference (biology-based models)
OECD 201	ISO 8692:2004 ISO 14593:1999 ISO 13829:2000	Alga growth inhibition	Growth rate	5.3	6.3	7.6
			Area under growth curve (biomass)	5.3	6.3	not recommended
OECD 202	ISO 6341:1996	Daphnia immobilization	Immobilization	5.2	6.2	7.3
OECD 203	ISO 7346-1:1996 ISO 7346-2:1996 ISO 7346-3:1996	Fish acute	Mortality	5.2	6.2	7.3
OECD 204		Fish prolonged	Mortality	5.2	6.2	7.3
			Body mass, length	5.3	6.3	7.4
		Avian acute	Mortality	5.2	6.2	7.3
OECD 205		Avian dietary	Mortality	5.2	6.2	7.3
			Body mass	5.3	6.3	7.4
			Food consumption	5.3	6.3	7.4 (theory covered, but not coded)
OECD 206		Avian-1-generation	Body mass (F0, F1), organ mass, food consumption, egg-shell thickness, egg-shell strength	5.3	6.3	7.4 (body mass)
			Egg production,	5.2/5.3	6.3	7.5
			14-day old survivors (counts)			
			Egg abnormality rate, egg fertility, viability, hatchability, chick survival rate (proportions)	5.2	6.2	7.3 (chick survival)
OECD 207	ISO 11268-1:1993	Earthworm acute	Mortality	5.2	6.2	7.3
			Body mass	5.3	6.3	7.4
OECD 208	ISO 11269-1:1993 ISO 11269-2:—[a]	Non-target terrestrial plant	Emergence	5.2	6.2	7.3 (as survival)
			Biomass, Root length	5.3	6.3	(theory covered, but not coded)
			Visual phytotoxicity	5.3	6.3	
			Mortality	5.2	6.2	7.3
	ISO 15522:1999	Activated sludge	Microorganism cell growth	5.3	6.3	7.6

Table 9 (*continued*)

Guideline/standard		Test	Endpoint	Reference (NOEC)	Reference (dose-response modelling)	Reference (biology-based models)
OECD 210		Fish ELS	Mortality	5.2	6.2	7.3
			Days to hatch	5.2/5.3	6.3	7.4 (theory covered, but not coded)
			Hatching success	5.2	6.2	7.3
			Days to swim-up	5.2/5.3	6.3	7.4 (theory covered, but not coded)
			Mass, length	5.3	6.3	7.4
OECD 211	ISO 10706:2000	Daphnia reproduction	Immobilization	5.2	6.2	7.3
			Fecundity	5.2/5.3	6.3	7.5
OECD 212	ISO 12890:1999	Fish embryo and sac-fry stage	Mortality	5.2	6.2	7.3
			Days to hatch	5.2/5.3	6.3	7.4 (theory covered, but not coded)
			Length	5.3	6.3	7.4
OECD 213		Honeybee, acute oral	Mortality	5.2	6.2	7.3
OECD 214		Honeybee, acute contact	Mortality	5.2	6.2	7.3
OECD 215	ISO 10229:1994	Fish juvenile growth test	Mortality	5.2	6.2	7.3
			Body mass, length	5.3	6.3	7.4
OECD 218/219		Chironomid toxicity	Emergence	5.2	6.2	7.3
			Days to hatch	5.2/5.3	6.3	7.4
			Survival	5.2	6.2	7.3
			Mass	5.3	6.3	7.4
OECD 220	ISO/CD	Enchytraeidae reproduction	Mortality	5.2	6.2	7.3
			Fecundity	5.2/5.3	6.3	7.4
	ISO 11268-2:1998	Earthworm reproduction	Mortality	5.2	6.2	7.3
			Body mass	5.3	6.3	7.4
			Fecundity	5.2/5.3	6.3	7.5
	ISO 11268-3:1999	Earthworm population size (field test)	Number of individuals (for various species)	5.2/5.3	6.3	7.6
OECD 221	ISO 20079:—[b]	Lemna growth inhibition	Average growth rate	5.3	6.3	7.4
			Area under growth curve	5.3	6.3	not recommended
			Final biomass	5.3	6.3	7.4
	ISO 10253:—[c]	Marine algal growth inhibition test	Growth rate	5.3	6.3	7.6
			Biomass	5.3	6.3	7.6
	ISO 14669:1999	Marine copepods	Immobilization	5.2	6.2	7.3
	ISO 11348-1:1998 ISO 11348-2:1998 ISO 11348-3:1998	Light emission of *Vibrio fischeri*	Luminescence	5.3	6.3	7.6 (for zero growth)

Table 9 (*continued*)

Guideline/standard		Test	Endpoint	Reference (NOEC)	Reference (dose-response modelling)	Reference (biology-based models)
	ISO 10712:1995	Pseudomonas putida growth inhibition	Growth rate	5.3	6.3	7.6
	ISO 13829:2000	Genotoxicity (umu-test)	Induction rate	5.3	6.3	
	ISO 11267:1999	Collembola reproduction inhibition	Offspring number	5.2/5.3	6.3	7.5
			Mortality	5.2	6.2	7.3

[a] To be published. (Revision of ISO 11269-2:1995)

[b] To be published.

[c] To be published. (Revision of ISO 10253:1995)

Annex A
(informative)

Analysis of an "acute immobilization of *Daphnia magna*" data set (OECD GL 202 — ISO 6341) using the three presented approaches

A.1 Data set (see Table A.1)

Table A.1 — Data set

Concentration	Number of immobilizations 0 h	Number of immobilizations 24 h	Number of immobilizations 48 h
0	0	0	0
0,39	0	0	0
0,78	0	0	0
1,56	0	0	1
3,13	0	0	7
6,25	0	6	11
12,5	0	5	14
25	0	5	20
50	0	19	20
100	0	20	20

A.2 Examples of data analysis using hypothesis testing (NOEC determination)

A.2.1 Example 1a — Daphnia acute example: Immobility after 48 h exposure

A.2.1.1 NOEC determination by two methods

The NOEC is 1,56 mg/l by both tests shown in the flow-chart, Figure 3. In this example, there was only a water control. Both Fisher's Exact test with a Bonferroni-Holm correction and the step-down Cochran-Armitage test are done to illustrate both approaches. In both cases, the overall false positive rate is controlled at 0,05 and one-sided tests are done comparing proportion alive in each concentration to that in the control. The p-values reported for Fisher's Exact test do not include the Bonferroni-Holm correction, but the determination of significance makes that adjustment.

A.2.1.2 Statistical analysis list file — Daphnia live 48 h

STATISTICAL ANALYSIS LIST FILE [22)]
DAPHNID LIVE48 OBSERVED DATA FROM DATASET ADAP_ACUTE
GROUP STATISTICS BY DOSE

dose	Number at risk	Number Responding	% Responding	Dose Score	Test Dose (mg/l)
1	20	20	100	1	0
2	20	20	100	2	0.39
3	20	20	100	3	0.78
4	20	19	95	4	1.56
5	20	13	65	5	3.13
6	20	9	45	6	6.25
7	20	6	30	7	12.5
8	20	0	0	8	25
9	20	0	0	9	50
10	20	0	0	10	100

A.2.1.3 Fisher's Exact test versus control

FISHER EXACT TEST vs CONTROL FOR DAPHNIA LIVE48 OBSERVED
TESTING FOR A DECREASING ALTERNATIVE HYPOTHESIS

Test Dose (mg/l)	Number at risk	Number Responding	Fisher's Exact Test *p*-value (Left)	Significance Rating
0.39	20	20	1.00000	
0.78	20	20	1.00000	
1.56	20	19	0.50000	
3.13	20	13	0.00416	*
6.25	20	9	0.00007	**
12.5	20	6	0.00000	**
25	20	0	0.00000	**
50	20	0	0.00000	**
100	20	0	0.00000	**

* refers to 0,01 < *p*-value < 0,05

** refers to a *p*-value < 0,01

A.2.1.4 Cochran-Armitage test using equally-spaced dose scores

Cochran-Armitage Test
Using Equally Spaced Dose Scores
Cochran-Armitage test is one-sided for DECREASE in RESPONSE
All doses included

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	141.11145	9	0	
Trend	-11.4263	1	1.545E-30	**
LOF	10.551094	8	0.2284543	

100 mg/l concentration omitted

22) In the annexes in this Technical Specification, the results of the tests shown in Courier New font include numbers with decimal points, instead of commas, as used elsewhere in this document.

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	119.23185	8	0	
Trend	-10.46522	1	6.239E-26	**
LOF	9.7109205	7	0.2055557	

100 and 50 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	93.867043	7	0	
Trend	-9.126828	1	3.527E-20	**
LOF	10.568053	6	0.1026794	

100, 50 and 25 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	58.680261	6	8.341E-11	
Trend	-7.068764	1	7.816E-13	**
LOF	8.7128292	5	0.1210814	

100, 50, 25 and 12.5 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	41.584158	5	7.1493E-8	
Trend	-5.637291	1	8.6373E-9	**
LOF	9.8051068	4	0.0438418	

100, 50, 25, 12.5 and 6.25 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	25.271739	4	0.0000444	
Trend	-3.909645	1	0.0000462	**
LOF	9.986413	3	0.018682	

100, 50, 25, 12.5, 6.25 and 3.13 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	3.0379747	3	0.3858072	
Trend	-1.350105	1	0.0884911	
LOF	1.2151899	2	0.5446592	

As discussed in Figure 3 of Clause 5 and the accompanying text, the Cochran-Armitage test is first applied to the entire data set. Given that that test is significant at the 0,05 level (*p*-value for trend is less than 0,05), the high concentration is omitted and the test is repeated with the remaining concentrations. This procedure is repeated until the Cochran-Armitage test is first not significant. The highest concentration remaining at that step is the NOEC. In the present case, that occurs when the highest remaining concentration is 1,56 mg/l, which is thus the NOEC.

Two other terms are shown at each stage. The first is labeled “Overall” and is the standard Chi-squared test with k-1 degrees of freedom for differences among the k groups represented. It is recalled that this Chi-squared statistic is the sum of the 1-d*f* Chi-squared statistic measuring linear trend (i.e. the square of the Cochran-Armitage test statistic) and the k-2 d*f* Chi-squared statistic for departure from linearity here labelled LOF.

A.2.2 Example 1b — Daphnia acute example: Immobility after 24 h exposure

A.2.2.1 NOEC determination by two methods

NOEC is 3,13 mg/l by the step-down Cochran-Armitage test and 25 mg/l by Fisher's Exact test. Both follow the flow-chart, Figure 3 of Clause 5. In this example, there was only a water control. Both Fisher's Exact test with a Bonferroni-Holm correction and the step-down Cochran-Armitage test are done to illustrate both approaches. In both cases, the overall false positive rate is controlled at 0,05 and tests are done comparing proportion alive in each concentration to that in the control. The p-values reported for Fisher's Exact test do not include the Bonferroni-Holm correction, but the determination of significance makes that adjustment

A.2.2.2 Statistical analysis list file — Daphnia live 24 h

STATISTICAL ANALYSIS LIST FILE
DAPHNID LIVE24 OBSERVED DATA FROM DATASET ADAP_ACUTE
GROUP STATISTICS BY DOSE

dose	Number at risk	Number Responding	% Responding	Dose Score	Test Dose (mg/l)
1	20	20	100	1	0
2	20	20	100	2	0.39
3	20	20	100	3	0.78
4	20	20	100	4	1.56
5	20	20	100	5	3.13
6	20	14	70	6	6.25
7	20	15	75	7	12.5
8	20	15	75	8	25
9	20	1	5	9	50
10	20	0	0	10	100

A.2.2.3 Fisher's Exact test versus control

FISHER EXACT TEST vs CONTROL FOR DAPHNIA LIVE24 OBSERVED
TESTING FOR A DECREASING ALTERNATIVE HYPOTHESIS

Test Dose (mg/l)	Number at risk	Number Responding	Fisher's Exact Test p-value (Left)	Significance Rating
0.39	20	20	1.00000	
0.78	20	20	1.00000	
1.56	20	20	1.00000	
3.13	20	20	1.00000	
6.25	20	14	0.01010	
12.5	20	15	0.02356	
25	20	15	0.02356	
50	20	1	0.00000	**
100	20	0	0.00000	**

A.2.2.4 Cochran-Armitage test using equally spaced dose scores

COCHRAN-ARMITAGE TEST
USING EQUALLY SPACED DOSE SCORES
Cochran-Armitage test is one-sided for DECREASE in RESPONSE

All doses included

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	136.55172	9	0	
Trend	-9.896625	1	2.153E-23	**
LOF	38.60853	8	5.8093E-6	

100 mg/l concentration omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	99.239409	8	0	
Trend	-7.804546	1	2.986E-15	**
LOF	38.328473	7	2.6241E-6	

100 and 50 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	30	7	0.000095	
Trend	-4.485426	1	3.6384E-6	**
LOF	9.8809524	6	0.1297557	

100, 50 and 25 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	30.190275	6	0.0000362	
Trend	-4.240398	1	0.0000112	**
LOF	12.209302	5	0.0320297	

100, 50, 25 and 12.5 mg/l concentrations omitted

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	31.578947	5	7.1979E-6	
Trend	-3.678836	1	0.0001172	**
LOF	18.045113	4	0.0012093	

When all concentrations at or above 6,25 mg/l are omitted, there is 100 % survival at every remaining concentration and no test can be done, nor is a test needed. The NOEC is 3,13 mg/l.

As discussed in Figure 3 of Clause 5 and the accompanying text, the Cochran-Armitage test is first applied to the entire data set. Given that that test is significant at the 0,05 level (p-value for trend is less than 0,05), the high concentration is omitted and the test is repeated with the remaining concentrations. This procedure is repeated until the Cochran-Armitage test is first not significant. The highest concentration remaining at that step is the NOEC. In the present case, that occurs when the highest remaining concentration is 1,56 mg/l, which is thus the NOEC.

Two other terms are shown at each stage. The first is labelled “Overall” and is the standard Chi-squared test with k-1 degrees of freedom for differences among the k groups represented. It is recalled that this Chi-squared statistic is the sum of the 1-df Chi-squared statistic measuring linear trend (i.e. the square of the Cochran-Armitage test statistic) and the k-2 df Chi-squared statistic for departure from linearity here labelled LOF.

A.3 Example of data analysis by dose-response modelling

A.3.1 General

Fitting a dose-response model to data is normally done with the aid of computer software. Therefore, the user does not need to worry about computational details. Outcomes from different software packages should be nearly identical when fitting the same model to the same data based on the same assumptions and methods. However, software packages may differ extensively in how to run them. In this subclause, the discussion of the examples is based on an analysis using the software package PROAST, but it is attempted to make the discussion helpful for users of other software just as well.

A.3.2 Bioassay on acute mortality of Daphnia magna

It is assumed that both an EC_{50} and an EC_{10} is required after two days of exposure. Therefore, a dose-response analysis of the mortality data at day 2 is normally sufficient. The following subclause discusses how such may be done.

A.3.3 Dose-response analysis for mortality at day 2

The data are quantal, and include more than one partial response. According to the flow-chart in Clause 6 (Figure 6), various models should be fitted in a way as described in 6.2.2. The first step is to put the data in the format that is required by the software to be used. Typically, the data shall be provided in a matrix form, as illustrated in Table A.2. Note that the information of the sample size in each concentration (dose) group is essential. Here, the model is fitted based on maximization of the log-likelihood assuming a binomial distribution for the observed responses.

Table A.2 — Data input file (as required by PROAST)

Daphnia_magna			
4			
1	4	0	0
Dose	Response	Sample_size	Day
100	20	20	1
50	19	20	1
25	5	20	1
12,5	5	20	1
6,25	6	20	1
3,13	0	20	1
1,56	0	20	1
0,78	0	20	1
0,39	0	20	1
0	0	20	1
100	20	20	2
50	20	20	2
25	20	20	2
12,5	14	20	2
6,25	11	20	2
3,13	7	20	2
1,56	1	20	2
0,78	0	20	2
0,39	0	20	2
0	0	20	2

A.3.4 Fitting the probit model

First, a probit model is fitted. As discussed in Clause 6, the background response should normally be estimated as a model parameter. However, when, as in this particular data set, fitting the model with or without a background parameter results in virtually the same outcome; the background parameter can just as well be omitted from the model (or, equivalently, be fixed at zero). It is preferable, in this case, to avoid non-convergence of the (iterative) fit algorithm. Note that in fitting the model without a background parameter, the data in the controls are not used, and they could just as well be deleted from the data set.

Figure A.1 shows the fit of the probit model for this data set. The parameter b is the EC_{50}, and the parameter c is the slope. It may also be interpreted as the inverse of the standard deviation (σ) of the underlying tolerance distribution (on the $\log_{10}$-scale).

Visual inspection of Figure A.1 indicates that the fitted model appears quite well supported by the data, and there is no reason to believe that the dose-response relationship could in fact be much different. To substantiate this, various other models are fitted, in concordance with the flow-chart in Clause 6.

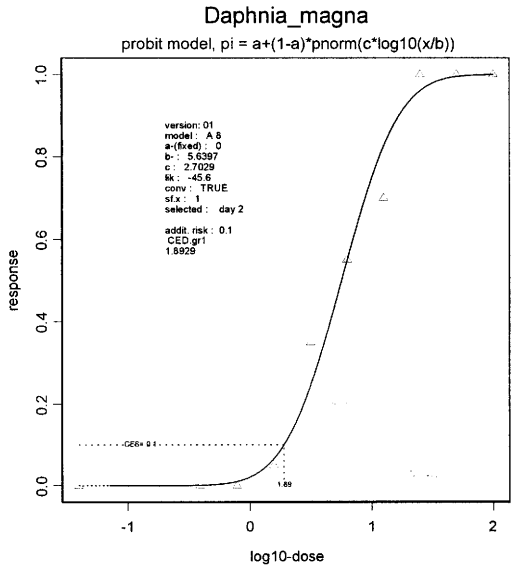


Figure A.1 — Probit model fitted to mortality response at day 2 — CED = EC_{10}

A.3.5 Fitting various models

Next to the probit model, three other models were fitted: the logit, the Weibull and the (two-stage) LMS model. As Table A.3 shows, the resulting EC_{50} and EC_{10} estimates are reasonably similar (less than 30 % difference). Also note that the EC_{10} is obtained by interpolation. Thus, it may be concluded that the data are suitable for deriving an EC_{10} by dose-response modelling.

Table A.3 — Summary of results regarding four models fitted to the mortality data in Figure A.1

Model	Log-lik	EC_{50}		EC_{10}	
		MLE [a]	90 %-CI [b]	MLE [a]	90 %-CI [b]
Probit	– 45,60	5,64	4,53 - 6,93	1,89	1,42 to 2,68
Logit	– 46,46	5,63	4,59 - 6,94	1,90	1,36 to 2,72
Weibull	– 46,25	6,64	5,36 - 8,17	1,69	1,10 to 2,68
Two-stage	– 46,83	7,00	5,59 - 8,46	1,47	1,17 to 1,77
[a] MLE = Maximum Likelihood Estimate.					
[b] CI = Confidence Interval; based on 1 000 bootstrap runs.					

A.3.6 Confidence intervals

The confidence intervals can be calculated by one of the ways described in Clause 6, depending on the software available. In Table A.2, the 90 %-confidence intervals are of the EC_{50} and the EC_{10} are given, as obtained by the bootstrap method (1 000 runs). They are not dramatically different between the models, and one may choose the lowest of these. Alternatively, one may choose the lower bounds associated with the probit model, as this model gives a somewhat higher log-likelihood value than the others. The slightly better fit can also be seen by comparing Figure A.1 with Figure A.2.

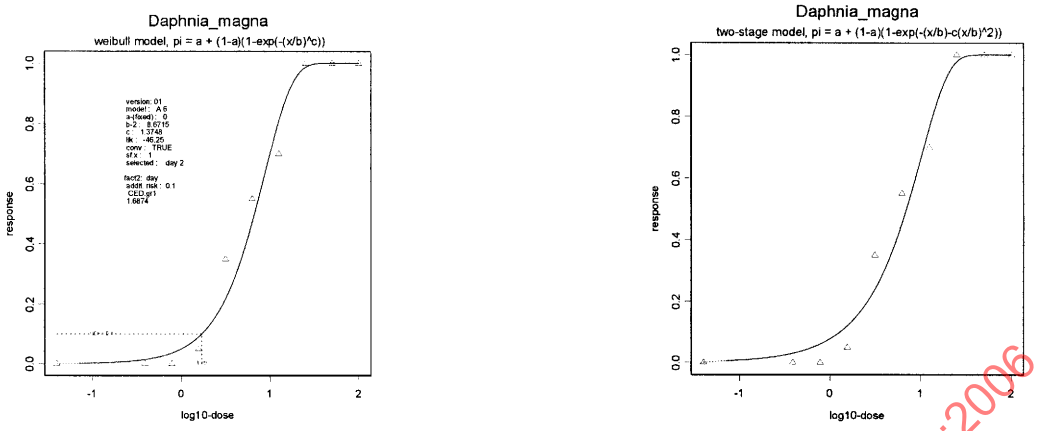


Figure A.2 — The Weibull (left panel) and the two-stage LMS model fitted to the mortality data at day 2

A.3.7 Mortality at both days

Although a single analysis of mortality at day 2 is normally sufficient, some remarks are made here on using the data at day 1 as well.

First, it may be noted that a similar analysis can be done for day 1 separately (see Figure A.3, left panel). Obviously, the EC_x values are higher, and for a risk assessment these results are probably not relevant. However, from the perspective of an efficient use of the data, an analysis in which the data from both days are analysed simultaneously would be preferable.

There are two ways of doing a simultaneous analysis.

— One is to fit the dose-response model to both days simultaneously, allowing for the fact that one (or some) of the parameters in the model depends on the number of days. This analysis is shown in the right panel of Figure A.3. Here, the parameter c (the slope) is assumed to be the same for both days, while the EC_{50} is allowed to differ between days. Thus, the simultaneous analysis estimated three parameters, while the two separate analyses together estimated four parameters in total.

  The sum of the log-likelihoods of the two separate analyses (−94,91) can be compared to the log-likelihood of the simultaneous analysis (−94,75). The former log-likelihood is associated with one more free parameter estimated from the data, and therefore can only be larger (or equal) to the latter log-likelihood. (NOTE: This holds in general for nested models, but not necessarily for non-nested models; two models are nested when one model can be derived from the other by reducing the number of parameters to be estimated from the data.)

— The likelihood ratio test can be used to assess if the larger number of parameters in the separate analyses is associated with in a significantly better fit. According to this test, the increase in log-likelihood should be at least 1,92 to be significant at $\alpha = 0{,}05$. It may therefore be concluded that the mortality data can just as well be described by a probit model with the same slope for days 1 and 2. As Table A.4 shows, the estimated EC_x values are quite similar between a simultaneous and a separate analysis.

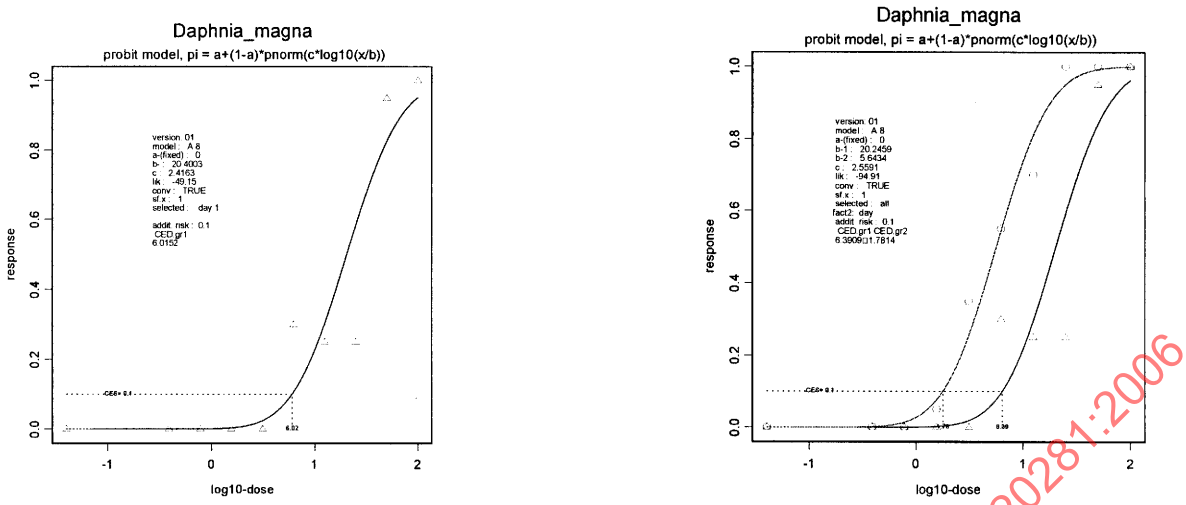


Figure A.3 — Probit model fitted to mortality data on day 1 (left panel), and fitted to both day 1 and day 2 simultaneously

Table A.4 — Results of fitting the probit model separately or simultaneously to day 1 and day 2

	Day 1		Day 2		Log-lik	Number of parameters estimated
	EC_{50}	EC_{10}	EC_{50}	EC_{10}		
Separate analysis	20,40	6,02	5,64	1,89	– 94,91 (n.s.)	4
Simultaneous analysis	20,25	6,39	5,64	1,78	– 94,75	3

The higher log-likelihood for the two separate analyses is not significantly higher than that for the simultaneous analysis.

The gain of a simultaneous analysis is twofold.

— First, one may expect that the estimated values of the EC_x are less likely to be biased (e.g. due to outliers or incidental systematic errors in the data), simply because they are based on more data points.

— Second, the data are described by less parameters, which complies with the parsimony principle (3.25).

One of the expected consequences is that the confidence intervals are smaller.

However, the analysis as now being discussed is not completely sound from a statistical point of view. The reason is that the data points for day 1 and day 2 are not independent, and therefore the confidence intervals may not be completely reliable. Nonetheless, they have been assessed for the EC_{50} and EC_{10} at day 2 (see Table A.5). They are wider than those obtained from a separate analysis of day 2, which is against expectation, since a fewer number of parameters was estimated in this analysis.

The reason is that in this particular case, the data at day 1 appear to show more noise than the data at day 2. Therefore, the noise is increased by including day 1 in the analysis, compared to an analysis of day 2 only. As long as there is no clear reason why the noise at day 1 may be larger than at day 2, it is hard to say if the noise for day 2 is in fact smaller, or only apparent or incidental. Thus, the increase in noise by including day 1 might be a more realistic reflection of the overall noise in the data.

Table A.5 — EC_{50} and EC_{10} at day 2, with 90 % confidence intervals, assessed by a simultaneous analysis of both days (see right panel of Figure A.3)

Model	EC_{50} at day 2		EC_{10} at day 2	
	MLE	90 %-CI [a,b]	MLE	90 %-CI
Probit	5,64	4,50 to 7,01	1,78	1,32 to 2,39

[a] It should be noted that these confidence intervals may not be completely reliable, due to dependencies in the data between day 1 and day 2.

[b] CI = Confidence Interval.

The dependencies in the data between day 1 and day 2 may be avoided by doing a simultaneous analysis the other way around. The mortality data may be regarded as a function of time for each dose group, while one of the parameters in the survival function is a function of dose. For instance, assuming a Weibull survival function, it might be assumed that the median survival time is some function of dose. Another approach is to assume a dose-response analysis for the hazard function (see, for example, the next subclause).

A.4 Example of data analysis using DEBtox (biological methods)

A.4.1 Parameters and asymptotic standard deviations (ASD)

(For a definition of all parameters, see Annex F.)

Survival, Hazard model			ASD	Correlation coefficients	
Blank mortality rate	1e-010	d^{-1}	0.000		
No-effect concentration-time	1.473	mg l^{-1} d	0.344	0.000	
Killing acceleration	0.07524	l mg^{-1} d^{-2}	0.010	0.000	0.359
Deviance	23.29				

Figure A.4 — DEBtox example: Parameters and asymptotic standard deviations (ASD)

A.4.2 Graphical test of model predictions against data

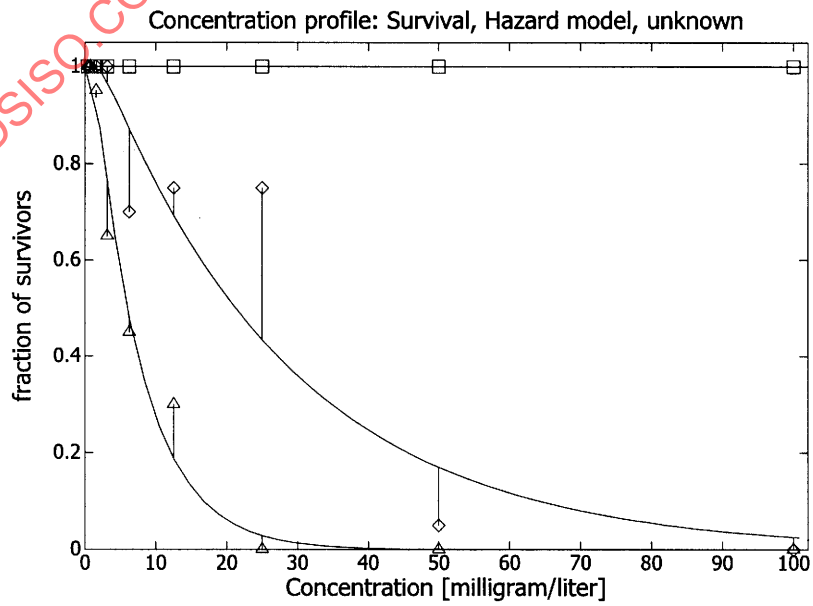


Figure A.5 — Graphical test of model predictions against data

Table A.6 — LC_x values (derived from parameter values) in milligrams/litre

Day	LC_0	ASD	LC_{50}	ASD
1	1,48	0,344	21,1	2,26
2	0,738	0,172	5,94	0,574

A.4.3 Comments

Slow kinetics appeared to fit the data best, which means that the elimination rate was too small to be estimated reliably. This means that the model loses this parameter. The consequence is that only the killing acceleration (which is the product of the killing rate and the elimination rate) can be estimated, not the killing rate itself. Similarly the ratio of the NEC and the elimination rate, called the no effect concentration time, could be estimated, rather than the NEC itself; the bioassay did not last long enough for this compound. The 95 % confidence interval for the NEC can still be estimated.

The background mortality rate was found to be nil. Notice that, excluding this parameter, a total of two parameters have been fitted on 18 data points.

Annex B
(informative)

Analysis of an "algae growth inhibition" data set using the three presented approaches

B.1 General

The following data set has been used. The aim of the presented analysis is to show the methodology that can be applied to these kinds of data.

Table B.1 — Data set on *Selenastrum capricornutum*

Concentration	Day 0	Day 0	Day 0	Day 1	Day 1	Day 1	Day 2	Day 2	Day 2
0	1,767	1,777	1,775	7,889	7,921	7,901	46,46	46,32	46,28
0	1,748	1,762	1,76	8,179	8,178	8,188	44,35	44,31	44,27
0	1,762	1,762	1,758	7,99	7,989	8,001	41,99	41,79	41,84
0	1,784	1,796	1,794	8,322	8,335	8,328	43,2	43,2	43,29
0	1,79	1,789	1,785	8,323	8,343	8,353	46,21	46,16	45,91
0	1,84	1,846	1,85	7,977	7,979	7,981	40,63	40,73	40,75
0,01	1,738	1,741	1,747	8,229	8,23	8,25	40,4	40,27	40,22
0,01	1,989	1,955	1,959	8,376	8,372	8,38	41,97	42,03	42,03
0,02	1,892	1,89	1,892	7,662	7,666	7,653	39,82	39,75	39,79
0,02	1,785	1,789	1,783	7,765	7,773	7,775	38,78	38,84	38,79
0,03	1,779	1,773	1,78	7,737	7,733	7,748	34,79	34,82	34,84
0,03	1,759	1,758	1,763	8,035	8,045	8,039	31,2	31,16	31,23
0,06	1,776	1,775	1,775	7,122	7,129	7,136	25,92	25,91	25,96
0,06	1,754	1,751	1,76	6,775	6,776	6,785	22,42	22,5	22,5
0,1	1,774	1,776	1,77	4,947	4,937	4,941	14,83	14,89	14,83
0,1	1,71	1,716	1,722	5,162	5,165	5,17	13,76	13,71	13,69
0,2	1,746	1,741	1,741	3,593	3,59	3,593	6,389	6,391	6,394
0,2	1,736	1,738	1,739	3,554	3,547	3,555	6,108	6,11	6,127
0,3	1,828	1,822	1,827	2,92	2,925	2,92	3,731	3,708	3,714
0,3	1,688	1,681	1,688	3,157	3,147	3,151	3,13	3,124	3,122
0,6	1,72	1,722	1,72	2,109	2,117	2,122	2,125	2,122	2,099
0,6	1,733	1,735	1,731	2,03	2,04	2,038	2,053	2,066	2,057

According to the July 2002, Draft Revised TG 201 [300], growth rate can be determined from cell count, biomass, or fluorescence values. That guideline refers to Mayer P., Cuhel R. and Nyholm N. (1997), a simple *in vitro* fluorescence method for biomass measurements in algal growth inhibition tests, *Water Research* **31**: pp. 2 525-2 531 on the use of fluorescence values for this purpose.

Indeed, any of these three measures of mass (cell count, biomass, or fluorescence) can be used to obtain very similar (but not necessarily identical) measures of growth rate on each replicate by fitting the model

$$y = a * e^{b*t},$$

using time, t, in hours, and y, the observed measure of mass. The estimate slope, b, from this fit is the sample growth rate for a given replicate. The simplest procedure is to linearize the problem by working with logarithms to obtain the model

$$\log(y) = A + b*t,$$

where $A = \log(a)$.

In what follows, natural logarithms were used, but this choice does not affect the slope or growth-rate estimate. These growth rate values can then be analysed by hypothesis testing or regression methods.

B.2 Examples of data analysis using hypothesis testing (NOEC determination)

B.2.1 Example 2a — Atrazine example: Fluorescence at Day 2

B.2.1.1 NOEC determination Atrazine by two methods

In Example 2a, the total biomass (or its surrogate, fluorescence) is analysed. In Example 2b, growth rate is analysed. NOEC is 0,01 mg/l by both tests. Both Dunnett's test and the step-down application of the Jonckheere test are illustrated, according to the chart in Figures 4 and 5 of 5.1.

B.2.1.2 Statistical analysis list file — Ecotoxicity measurements from data set Atrazine

STATISTICAL ANALYSIS LIST FILE
ECOTOX MEASUREMENTS FROM DATASET ATRAZINE
GROUP STATISTICS FOR Average_2 BY DOSE

dose	doseval	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	6	43.5278	43.5373	2.26521	0.92477
2	0.01	2	40.9206	40.9206	1.21151	0.85667
3	0.02	2	39.0623	39.0623	0.69532	0.49167
4	0.03	2	32.7739	32.7739	2.55973	1.81000
5	0.06	2	23.9689	23.9689	2.44423	1.72833
6	0.1	2	14.0523	14.0523	0.79903	0.56500
7	0.2	2	6.0204	6.0204	0.19540	0.13817
8	0.3	2	3.1888	3.1888	0.41884	0.29617
9	0.6	2	1.8543	1.8543	0.04007	0.02833

B.2.1.3 Shapiro-Wilk test of normality of average 2

SHAPIRO-WILK TEST OF NORMALITY OF Average_2

STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
1.39706	-0.099030	0.10954	0.97324	0.78479	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as sole factor. The Shapiro-Wilk test does not indicate a problem with the normality assumption. The Tukey outlier rule is used to identify observations that may be of special interest. Outliers can have an impact on the results, as well as on the assessment of normality and variance homogeneity. A possibly valuable use of these outliers would be to re-run the analysis with outliers omitted to determine whether the NOEC is affected by these outliers. If it is, then caution should be followed in using the results. It is important to understand that just because an observation is identified as an outlier, that does not mean the observation is "bad" or that it should

not be used. An observation should be excluded from analysis only for scientifically sound reasons and any such exclusion shall be clearly stated along with its justification.

B.2.1.4 Outliers and influential observations

Outliers and Influential Observations

SELENASTRUM	Dose	Doseval	Group	OBSER	Pred	Resid
1	1	0	I	46.1206	43.5278	2.59278
5	1	0	I	45.8606	43.5278	2.33278
6	1	0	I	40.4706	43.5278	-3.05722

B.2.1.5 Levene test for average 2

LEVENE TEST FOR Average_2 - FULL Model

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	8	3.36022	0.025754	**

The data was found to be normally distributed (the p-value for the SW test was 0,784 79) but group variances were unequal (the p-value for Levene's test was 0,025 754). A Tamhane-Dunnett analysis is appropriate.

B.2.1.6 Tamhane-Dunnett one-sided test for decrease in means in average 2

Tamhane-Dunnett one-sided test for decrease in means in Average_2
Using MAXIMUM LIKELIHOOD estimates of variation on ECOTOX values.

dose	MEAN	STDERR	degfree	CONTROL	OBS_DIFF	crit	SIGNIF
2	40.9206	0.85667	3.68718	43.5278	-2.6072	5.3026	
3	39.0623	0.49167	5.87792	43.5278	-4.4656	3.5625	*
4	32.7739	1.81000	1.56884	43.5278	-10.7539	13.7828	
5	23.9689	1.72833	1.62787	43.5278	-19.5589	13.2920	*
6	14.0523	0.56500	5.56760	43.5278	-29.4756	3.7643	*
7	6.0204	0.13817	5.21273	43.5278	-37.5074	3.3204	*
8	3.1888	0.29617	5.77453	43.5278	-40.3391	3.3255	*
9	1.8543	0.02883	5.00937	43.5278	-41.6736	3.3279	*

Thus, the Tamhane-Dunnett test finds significant decreases in mean response at dose 3 = 0,02 mg/l and at 0,06 mg/l and above, but not at dose 4 = 0,03 mg/l. So, there is some departure from monotonicity in the dose response. Whether the NOEC should be set at 0,03 or at 0,01 from this test is a matter of scientific judgement.

B.2.1.7 Monotonicity check of average 2

MONOTONICITY CHECK OF Average_2 - FULL DATA
DOSES 0, 0.01, 0.02, 0.03, 0.06, 0.1, 0.2, 0.3, 0.6 mg/l

PARM	P_T	SIGNIF
DOSE TREND	0.00000	**
DOSE QUAD	0.83911	

This is a formal test for departure from monotonicity in the dose response. It is based on linear contrasts, as described in 5.1.4. In this instance, there is evidence of a linear dose-response response (p-value for dose trend < 0,000 01), so there is no reason, based on this test, to doubt the overall monotonicity of the dose response. We accordingly proceed with the step-down Jonckheere-Terpstra test.

In the results presented below,

— JONC is the value of the Jonckheere-Terpstra test statistics;

— P1DNCF is the p-value associated with this test statistic to test the hypothesis of a downward trend;

— P1UPCF is the test statistic for testing the significance of an upward trend and is a default printout of the software used and is not utilized in the present analysis, i.e. the p-value for the upward trend;

— ZC is a standardized value of the JONC statistic, computed with tie correction;

— ZCCF is ZC, except that it is computed using a standard continuity correction factor.

The CF in several terms refers to the use of this continuity correction factor.

p-values are for the tie-corrected test with the continuity correction factor.

SIGNIF RESULTS are for a decreasing alternative hypothesis.

```
Jonckheere Trend Test on Dose 0 + Lowest 8 Doses through 0.6 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -5.837314   -5.8087         1 2.2331E-9 **

Jonckheere Trend Test on Dose 0 + Lowest 7 Doses through 0.3 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -5.422661 -5.389596         1 2.4387E-8 **

Jonckheere Trend Test on Dose 0 + Lowest 6 Doses through 0.2 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -4.972358 -4.933512 0.9999996 2.7045E-7 **

Jonckheere Trend Test on Dose 0 + Lowest 5 Doses through 0.1 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -4.476023 -4.429398 0.9999953 3.0535E-6 **

Jonckheere Trend Test on Dose 0 + Lowest 4 Doses through 0.06 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -3.917286 -3.859679 0.9999432 0.0000352 **

Jonckheere Trend Test on Dose 0 + Lowest 3 Doses through 0.03 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -3.267487 -3.193226 0.9992965 0.0004163 **

Jonckheere Trend Test on Dose 0 + Lowest 2 Doses through 0.02 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -2.466679 -2.363901 0.9909582 0.0050929 **

Jonckheere Trend Test on Dose 0 + Lowest 1 Doses through 0,01 mg/l

             JONC        ZC      ZCCF    P1UPCF    P1DNCF SIGNIF
                2 -1.333333 -1.166667 0.8783275 0.0668072
```

The Jonckheere-Terpstra test is significant at the 0,05 level at each step until the highest remaining concentration used in evaluating that statistics is 0,01 mg/l. Accordingly, the NOEC is set at 0,01 mg/l. Group means should be examined to check for lack-of-fit to a linear trend before trend test results are accepted. That is, blind reliance on the formal test for monotonicity is not advised. Rather, an informed judgement should be made based on all the available information. The same holds for assessing normality.

B.2.2 Example 2b — Atrazine example: Growth rate

B.2.2.1 NOEC determination Atrazine by two methods

In this subclause, an NOEC is obtained for growth rate, both by Dunnett's test and by the step-down application of the Jonckheere-Terpstra test. For ease of reference, a table of growth rates obtained from the fluorescence data is given below.

Conc	REPLICATE	G_Rate	Conc	REPLICATE	G_Rate
0	A	0,07082	0,03	B	0,06269
0	B	0,07010	0,06	A	0,05860
0	C	0,06886	0,06	B	0,05587
0	D	0,06911	0,1	A	0,04688
0	E	0,07050	0,1	B	0,04599
0	F	0,06714	0,2	A	0,02929
0,01	A	0,06831	0,2	B	0,02840
0,01	B	0,06628	0,3	A	0,01631
0,02	A	0,06608	0,3	B	0,01435
0,02	B	0,06692	0,6	A	0,00490
0,03	A	0,06476	0,6	B	0,00409

The NOEC is 0,01 mg/l by the step-down Jonckheere-Terpstra test, as well as by Dunnett's test.

B.2.2.2 Statistical analysis list file — Ecotox measurements from data set ATRZ_GRATES

STATISTICAL ANALYSIS LIST FILE
ECOTOX MEASUREMENTS FROM DATASET ATRZ_GRATES
GROUP STATISTICS FOR Estimate BY DOSE

dose	Conc	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	6	0.069421	0.069604	.001355632	.000553435
2	0.01	2	0.067294	0.067294	.001435844	.001015295
3	0.02	2	0.066500	0.066500	.000598875	.000423469
4	0.03	2	0.063729	0.063729	.001462545	.001034175
5	0.06	2	0.057236	0.057236	.001932705	.001366629
6	0.1	2	0.046432	0.046432	.000626564	.000443048
7	0.2	2	0.028843	0.028843	.000627415	.000443649
8	0.3	2	0.015327	0.015327	.001389178	.000982297
9	0.6	2	0.004497	0.004497	.000571829	.000404344

B.2.2.3 Shapiro-Wilk test of normality of estimate

SHAPIRO-WILK TEST OF NORMALITY OF Estimate

OBS	STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
22	.000988660	-0.42123	-0.42443	0.94515	0.25242	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as sole factor. The non-significant p-value (p = 0,25242) for the Shapiro-Wilk test indicates no reason to reject the normality assumption. The Tukey outlier rule identified no observations of special interest and hence, no outliers are reported.

B.2.2.4 Levene test for estimate — Full model

LEVENE TEST FOR Estimate - FULL Model

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	8	1.38466	0.28901	

The data was found to be consistent with a normal distribution, as shown by the Shapiro-Wilk test above, and with equal variances, as shown by the Levene test immediately above. An analysis of variance is performed.

Obs	Class	Levels	Values
1	dose	9	1 2 3 4 5 6 7 8 9

B.2.2.5 Overall F-tests for ANOVA

OVERALL F-TESTS FOR ANOVA

Obs	Effect	Num DF	Den DF	FValue	ProbF
1	dose	8	13	897.85	<.0001

As discussed in Clause 5, no use is made of this result, though it does indicate that there is significant variation among the treatment means. It is a default computer output. The reader shall be prepared to make intelligent use of default output.

B.2.2.6 Estimated dose effects and Dunnett for decreasing alternatives using Alpha = 0,05

ESTIMATED DOSE EFFECTS & DUNNETT FOR Decreasing ALTERNATIVE
USING ALPHA=.05 FOR COMPARISONS TO CONTROL

Estimate	SIGNIF	Dunnett one-sided p-value	Test Group Mean	N
DOSE TREND	**	0.00000	.	.
DOSE QUAD	**	0.00000	.	.
DOSE 2-1		0.16679	0.067294	2
DOSE 3-1	*	0.04470	0.066500	2
DOSE 4-1	**	0.00035	0.063729	2
DOSE 5-1	**	0.00000	0.057236	2
DOSE 6-1	**	0.00000	0.046432	2
DOSE 7-1	**	0.00000	0.028843	2
DOSE 8-1	**	0.00000	0.015327	2
DOSE 9-1	**	0.00000	0.004497	2

The significant Dose Trend result indicates that there is a significant linear trend in the dose response and no reason, by this formal test, to question the monotonicity of the dose response. That there is also a significant quadratic trend is generally an indication that the overall trend is not linear. It does not, in itself, indicate non-monotonicity. The plots in the regression analysis of these data are instructive in this regard. An inspection of the treatment means reveals that the means are indeed monotone.

The Jonckheere-Terpstra test does not require linearity, only monotonicity.

By Dunnett's test, the NOEC is 0,01 mg/l, the lowest concentration.

B.2.2.7 Check for ties in estimate

```
Check for ties in Estimate
Percent of all data tied at 3 most frequently observed values
Since 5 < 25%, 9 < 40% and 14 < 65%
Exact methods (StatXact) are not required on this basis.
```

COUNT	SUMWTS	NMISS	NOBS	RESPONSE	TIES	TIEPCT
1	22	2	24	0.004093	1	5
1	22	2	24	0.004901	2	9
1	22	2	24	0.014345	3	14

B.2.2.8 Jonckheere-Terpstra test — Concentrations through groups 2, 3, 4, 5, 6, 7, 8 and 9

The small number of replicates per concentration, two (except in the control) suggests an exact Jonckheere-Terpstra test may be warranted. Certainly, such a test is not wrong. Below, both the exact and asymptotic (large-sample) results are given. It is observed that they do not agree. The exact result is considered definitive and is used to declare the NOEC to be 0,01 mg/l, the same as by Dunnett's test for this example. As is seen below, each Jonckheere-Terpstra test is significant down to the 0,02 mg/l concentration. The result for the final test, where only the lowest concentration and control remain, differ in the exact and asymptotic versions of the test.

```
      Concentrations through Group 9

       Jonckheere-Terpstra Test
-------------------------------------
Statistic (JT)                 13.5000
Z                              -5.8373
Asymptotic Test
One-sided Pr <  Z               <.0001
Exact Test
One-sided Pr <= JT           8.691E-15

      Concentrations through Group 8

       Jonckheere-Terpstra Test
-------------------------------------
Statistic (JT)                 13.0000
Z                              -5.4227
Asymptotic Test
One-sided Pr <  Z               <.0001
Exact Test
One-sided Pr <=  JT          1.629E-12

      Concentrations through Group 7

       Jonckheere-Terpstra Test
-------------------------------------
Statistic (JT)                 12.5000
Z                              -4.9724
Asymptotic Test
One-sided Pr <  Z               <.0001
Exact Test
One-sided Pr <=  JT          2.447E-10
```

Concentrations through Group 6

Jonckheere-Terpstra Test

Statistic (JT)	12.0000
Z	-4.4760
Asymptotic Test	
One-sided Pr < Z	<.0001
Exact Test	
One-sided Pr <= JT	2.863E-08

Concentrations through Group 5

Jonckheere-Terpstra Test

Statistic (JT)	11.5000
Z	-3.9173
Asymptotic Test	
One-sided Pr < Z	<.0001
Exact Test	
One-sided Pr <= JT	2.511E-06

Concentrations through Group 4

Jonckheere-Terpstra Test

Statistic (JT)	11.0000
Z	-3.2675
Asymptotic Test	
One-sided Pr < Z	0.0005
Exact Test	
One-sided Pr <= JT	1.563E-04

Concentrations through Group 3

Jonckheere-Terpstra Test

Statistic (JT)	10.5000
Z	-2.4667
Asymptotic Test	
One-sided Pr < Z	0.0068
Exact Test	
One-sided Pr <= JT	0.0063

Concentrations through Group 2

Jonckheere-Terpstra Test

Statistic (JT)	9.0000
Z	-1.6667
Asymptotic Test	
One-sided Pr < Z	0.0478
Exact Test	
One-sided Pr <= JT	0.0714

B.3 Example of data analysis by dose-response modelling

B.3.1 Data set and approach

It is assumed that an EC_{10} is required.

The data consist of observed biomass at three consecutive days, where exposure starts at the first day (= day 0). Each flask is sampled at the three points in time, so the data are not independent in time.

One approach for dose-response analysis of this type of data is to fit a dose-response model to the biomass observations for day 1 and day 2 separately. In practice, the analysis of one day only is actually used (the one that results in the lowest EC_{10}).

Another approach that does make use of all data available is to consider not the biomass but the growth rate as a function of concentration. From a statistical point of view, this approach is more efficient (in using all the data in a single analysis). Further, one may argue that changes in biomass result from changes in growth rate, so that growth rate is a more relevant parameter from a biological point of view. Unfortunately, no consensus exists on this issue. This second approach is illustrated for the Atrazine data set (this analysis is also briefly discussed in Clause 6).

B.3.2 Observations

Figure B.1 shows the observed biomass (i.e. fluorescence) data at days 0, 1 and 2, where nine different concentrations (including zero) of Atrazine have been applied. An exponential model,

$y = a \exp(bx)$, was fitted to the data,

where

a denotes the initial biomass at day zero; and

b the growth rate.

Given the experimental protocol, the initial biomass cannot depend on the concentration, and therefore the parameter a in this growth model can be assumed constant. The parameter b is estimated by allowing it to be dependent on the concentration, i.e. for each concentration a particular value for b is estimated. Thus, a total of 11 parameters are estimated from this dataset: one value for a, nine values for b, and one value for var, the residual variance (the variance of the residuals[23)] on the log-scale).

B.3.3 Analysis

B.3.3.1 First step

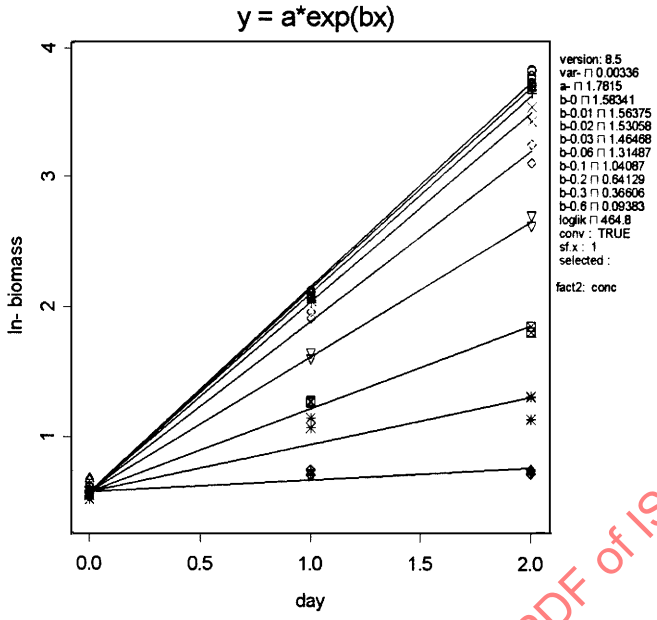
In the analysis underlying Figure B.1, the observations are assumed to be independent. However, as discussed above, at each concentration various flasks were used, and each flask was sampled at the three days of observation, i.e. the individual flasks were followed in time. Therefore, the data as plotted in Figure B.1 are not independent, while nonetheless independence was assumed in fitting the model.

This problem may be circumvented by estimating a growth rate for each individual flask. Since 22 flasks were observed (6 in the control, and 2 in the other concentration groups), a total of 24 parameters is estimated in such an analysis (including a and var). Figure B.2 shows the results of this analysis.

The log-likelihood has increased from 464,80 (see Figure B.1) to 483,86 (see Figure B.2). This increase of 19,06 log-likelihood units for a model with 24 - 11 = 13 more parameters is highly significant according to the

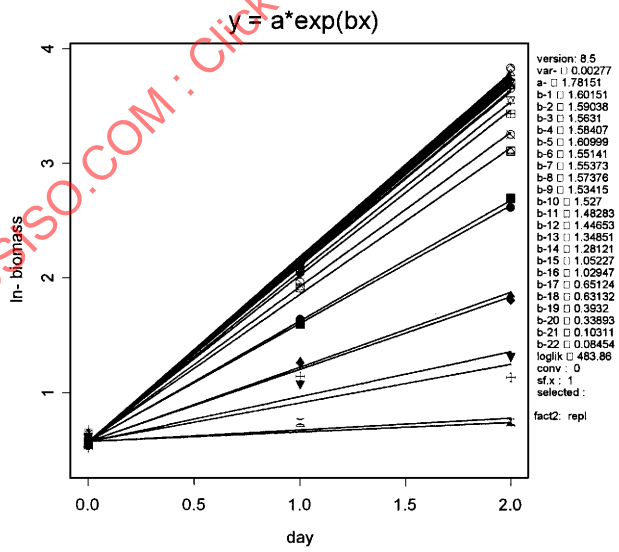
23) The residuals are the deviations of the individual data points from the fitted model.

likelihood ratio test (twice the difference in log-likelihood is approximately Chi-squared distributed with 13 degrees of freedom; $P \approx 0{,}000\ 3$). Hence, it may be concluded that individual flasks differ from each other by themselves.



NOTE Biomass is plotted on the log-scale, resulting in linear growth curves. Biomass was assumed to be log-normally distributed, with homogenous variance on log-scale (i.e. homogenous CV). Here, 11 parameters are estimated (in a simultaneous fit).

Figure B.1 — Exponential growth model fitted to biomass, assuming a constant initial biomass (a), and growth rate (b) dependent on concentration (0, 0,01, 0,02, 0,03, 0,06, 0,1, 0,2, 0,3, 0,6 mg/l)



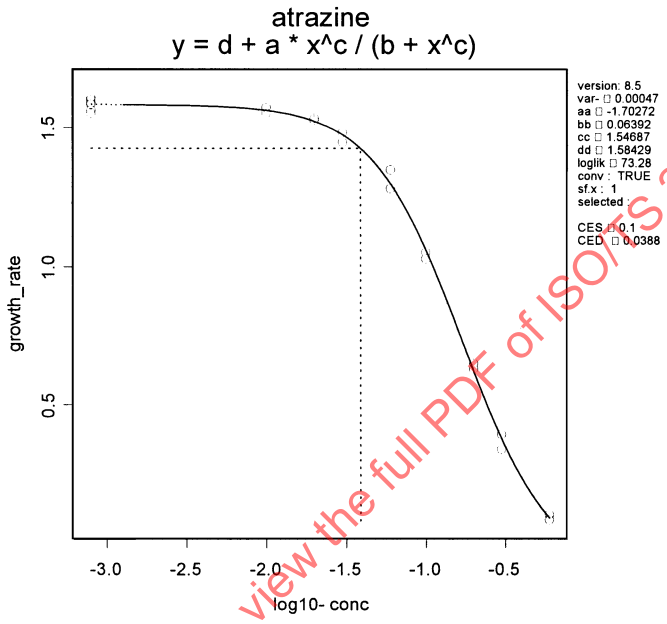
NOTE A total of 24 parameters are estimated (in a simultaneous fit). By comparing the log-likelihood (483,86) with that obtained in Figure B.1 (loglik = 464,80), it may be concluded that a significantly better fit is obtained, implying that flasks (at the same concentrations) are different with respect to growth rates.

Figure B.2 — Exponential growth model fitted to biomass, assuming a constant initial biomass (a), and growth rate (b) dependent on each individual flask (six for the control group, and 2 for each nonzero concentration)

B.3.3.2 Second step

As a second step in the analysis, the estimated growth rates are plotted against the concentration.

Figure B.3 shows the growth rates from Figure B.2, as a function of concentration, and a dose-response model may be fitted to these data. Here, it is assumed that the growth rates are normally rather than log-normally distributed (one of the reasons being that negative growth rates are possible). Figure B.3 shows the results for the Hill model fitted to the growth rates. This model fits the data extremely well, while the resulting curve is well confined by the data. Thus, estimation of an EC_{10} is fully warranted. It is no surprise that different models give very similar results and very narrow confidence intervals (see Table B.2).



NOTE The Hill model is fitted here to the data. The data are plotted against log-concentration to improve visibility.

Figure B.3 — Estimated growth rates (from individual flasks, see Figure B.2) as a function of the concentration of Atrazine

Table B.2 — Summary of results of dose-response analysis for growth rate

Model	Log-lik	EC_{10}	90 % CI
$y = d + a\,x^c / (b^c + x^c)$	73,28	0,039	0,035 5 to 0,042 1 [a]
$y = a\,[c - (c\text{-}1)\exp(-x/b)^d]$	71,17	0,035	0,032 2 to 0,038 6 [b]

[a] Based on 5 000 bootstrap runs.

[b] Based on likelihood profile method.

B.3.4 Assumptions

B.3.4.1 General

To check the assumptions of normality and homogeneous variances, the regression residuals (i.e. the deviations of the individual data points from the dose-response model) may be plotted in various ways, e.g. the so-called QQ-plot. In a QQ-plot, the observed quantiles are plotted against the theoretical quantiles, e.g. according to the normal distribution. When data are sampled from a normal distribution, this plot should theoretically result in a straight line. It should be noticed that fitting a line to a QQ-plot is unsound (which is not always recognized). One may draw the theoretical straight line in the plot, with intercept equal to the mean of

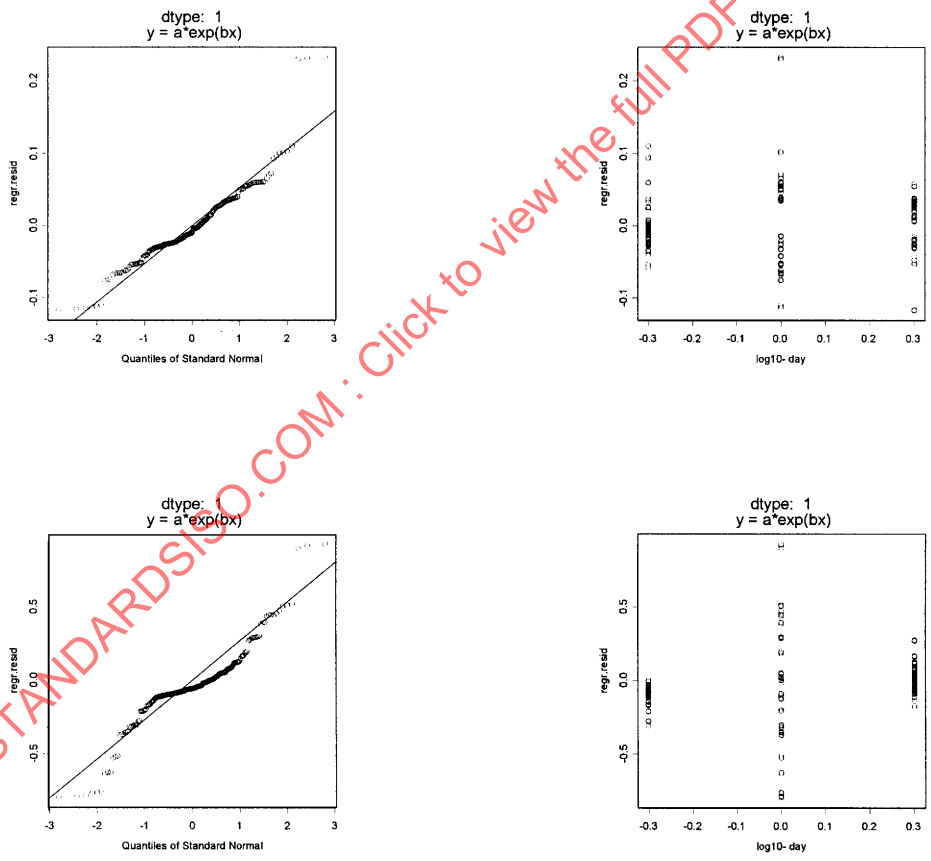
the data points and with slope equal to the standard deviation of the data points. In the case of regression, the data points are the regression residuals, which are corrected for the dose-response relationship.

In interpreting a QQ-plot, one should realize that, due to sampling errors, fluctuations around the line can easily arise, especially in small data sets. In particular, a pattern resembling Aesculapius' staff is not unusual, even for data that are sampled from a normal distribution by the computer. Hence, QQ-plots should only lead to the conclusion that the assumed distribution is inadequate when the data show a clear overall curvature. It is always the general trend, not single data points that should be considered.

B.3.4.2 Biomass data

The biomass observations were assumed to be log-normally distributed, and therefore the model was fitted on log-scale. Hence, the residuals on a log-scale should be normally distributed with zero mean. From the left panel of Figure B.4 it may be concluded that the assumption of log-normally distributed biomass is reasonable. In the right panel of Figure B.4 the regression residuals are plotted for the three days separately. This plot reveals no differences in scatter between days, and it may be concluded that the assumption of homogeneous variances (on log-scale) is reasonable as well.

For the sake of illustration, the same plots are shown for the residuals resulting from an analysis without transformation. The QQ-plot shows a much less linear relationship, while the scatter is clearly not homogeneous between days. Clearly, an analysis after logarithmic transformation is more adequate.

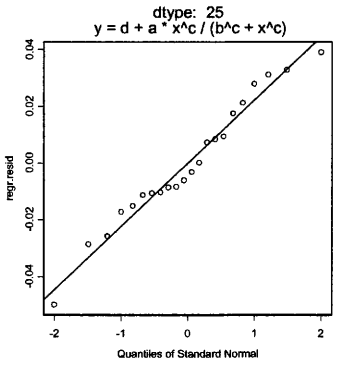


NOTE The QQ-plots (left panels) show that the data comply with the assumption of log-normality, but less so with normality. The variances appear to be homogeneous for the analysis on log-scale (upper right panel), with some outliers at day 1 (middle group). The same three outliers are visible in the QQ-plot (upper left panel). The analysis without transformation [lower right panel] results in a large scatter in the residuals at day 1 (middle group), and the assumption of homogeneous variances is clearly violated.

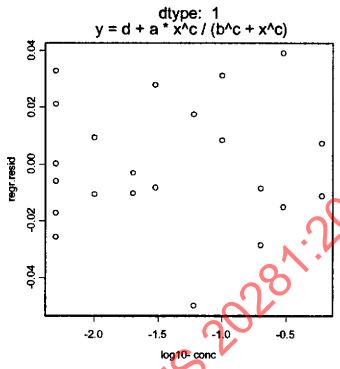
Figure B.4 — Regression residuals from analysis on log-scale (upper panels) and from analysis without transformation (lower panels)

B.3.4.3 Growth rates

The growth rates were analysed without transformation, i.e. they were assumed to be normally distributed themselves, with homogeneous variance among concentration groups. As Figure B.5 shows, both assumptions appear reasonable.



a) QQ-plot for regression residuals for growth rates, confirming the assumption of normality



b) Regression residuals plotted against (log-)concentration, confirming the assumption of homogeneous variances

Figure B.5 — Growth rate plots

B.3.5 Dependencies due to individual flasks

The dose-response analysis discussed here was based on the growth rates derived for the individual flasks. As already discussed in Clause 6, an analysis that is based on an estimated growth rate for each concentration group (as in Figure B.1) results in virtually the same estimate for the EC_{10} and for its confidence interval. Yet, the first analysis (separate growth rate estimate for each flask) is favourable, as it may give the information that a particular flask might be an outlier. Further, it may give information on weaknesses in the study protocol, for instance when replicate flasks in the same concentration group deviate from the general dose-response pattern. Such would indicate that the experimental protocol could be improved by better randomization. In this way, the test could be made more effective.

B.4 Examples of data analysis using DEBtox (biological methods)

B.4.1 Data

Figure B.6 shows the data set used to analyse the effects of Atrazine in micrograms per litre (µg/l) on the growth of *Selenastrum capricornutum* in cells per millilitre (cells/ml).

Time: hour, Conc: microgram/liter, Resp: cells ml-1																					
	0	0	0	0	0	0	10	10	20	20	30	30	60	60	100	100	200	200	300	300	600	600
0	1.5403	1.5239	1.5279	1.5386	1.5353	1.6126	1.5093	1.7349	1.6386	1.5529	1.5446	1.5273	1.5426	1.5223	1.5406	1.4833	1.5099	1.5049	1.5929	1.4529	1.4879	1.5003
1	7.6709	7.9489	7.7606	8.0956	8.1069	7.7463	8.0036	8.1433	7.4276	7.5383	7.5066	7.8069	6.8963	6.5459	4.7089	4.9329	3.3593	3.3193	2.6889	2.9189	1.8833	1.8033
2	46.1206	44.0773	41.6406	42.9973	45.8606	40.4706	40.0639	41.7773	39.5539	38.5706	34.5839	30.9639	25.6973	22.2406	14.6173	13.4873	6.1586	5.8823	3.4849	2.8926	1.8826	1.8259

Figure B.6 — DEBtox example: Data for effects of Atrazine in micrograms per litre on the growth of *Selenastrum capricornutum* in cells per millilitre

Figure B.7 shows the parameter estimates and Asymptotic Standard Deviations (ASD).

Population growth, Growth model			ASD	Correlation coefficients		
Inoculum size	1.446	·cells ml^{-1}	0.099			
Population growth rate	1.695	h^{-1}	0.035	-0.991		
No-effect concentration	15.61	ug l^{-1}	1.160	0.106	-0.161	
Tolerance concentration	176.6	ug l^{-1}	10.834	-0.570	0.559	-0.489
Mean deviation	1.13	·cells ml^{-1}				

Figure B.7 — DEBtox example: Parameter estimates and asymptotic standard deviations (ASD)

B.4.2 Graphical test

Figures B.8 and B.9 show graphical tests of model predictions against data (the different curves correspond to the different concentrations).

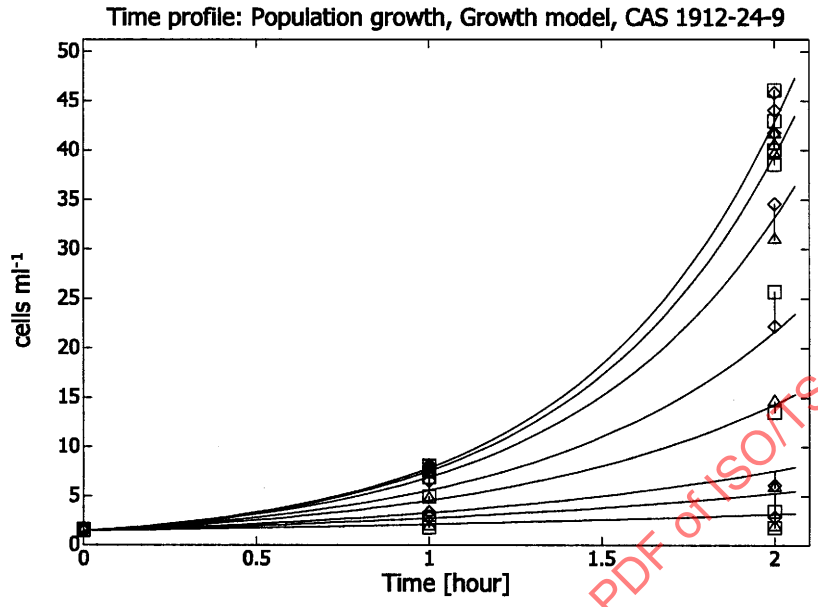


Figure B.8 — DEBtox example: Time profile (Population growth, growth model, CAS 1912-24-9)

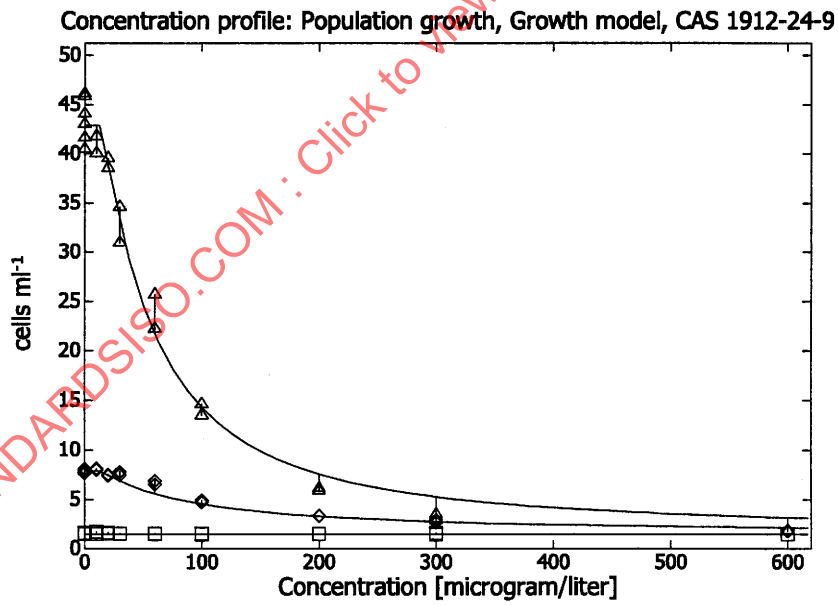


Figure B.9 — DEBtox example: Concentration profile (Population growth, growth model, CAS 1912-24-9)

B.4.3 Profile likelihood for NEC estimate

First select the confidence level of your choice in the left panel (see Figure B.10), then read the ln likelihood. The concentrations in the right panel for which the ln likelihoods are below this level comprise the confidence set of the NEC.

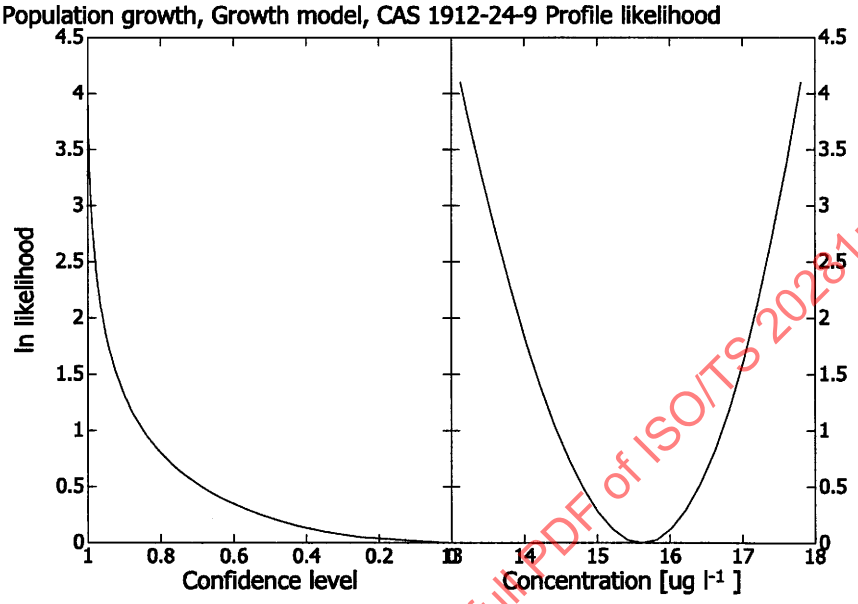


Figure B.10 — DEBtox example: Profile likelihood for NEC estimate (Population growth, growth model, CAS 1912-24-9)

Table B.3 — EC_x values

Day	EC_0	ASD	EC_{50}	ASD
1	15,6	1,16	139	5,75
2	15,6	1,16	61	2,03

Comments

The model for effects on the growth rate fits quite acceptably, but those for effects on adaptation and hazard fitted slightly better with similar NEC values (see Table B.4). The effects on growth have been selected here to improve the comparability with the concentration-response method. The 99 % confidence intervals for the NEC values in micrograms per litre (μg/l) for the three models are:

Table B.4 — 99 % confidence intervals for the NEC values for the three models

Model	99 % Confidence intervals	
Hazard	5,89	14,4
Adaptation	4,46	9,59
Growth	13,2	17,4

These values are obtained by selecting the different models in the DEBtox software, and using its routine for computation of the confidence interval of the NEC.

Annex C
(informative)

Analysis of an "*Daphnia magna* reproduction" data set (OECD GL 211 – ISO 10706) using the three presented approaches

Table C.1 — Data set on *Daphnia magna* — Data for the cumulative number of offspring per female as affected by an unknown compound

Concentration mg/l	Day	Number of live young produced						
		Rep A	Rep B	Rep C	Rep D	Rep E	Rep F	Rep G
0	7	0	0	0	0	0	0	0
0	10	13	4	12	0	8	7	6
0	12	2	14	0	0	0	0	0
0	14	13	0	14	14	13	14	34
0	17	25	27	31	19	32	24	38
0	19	34	34	40	24	40	18	38
0	21	0	0	0	0	0	0	0
0,014	7	0	0	0	0	0	0	0
0,014	10	8	6	12	5	8	12	7
0,014	12	2	0	0	0	2	0	0
0,014	14	20	16	23	14	14	20	15
0,014	17	27	24	26	27	25	32	32
0,014	19	34	27	41	33	24	35	35
0,014	21	0	0	0	0	0	0	0
0,050	7	0	0	0	0	0	0	0
0,050	10	10	12	11	10	1	0	0
0,050	12	0	0	0	0	18	2	0
0,050	14	12	21	22	17	0	3	11
0,050	17	26	36	26	30	8	21	0
0,050	19	37	29	28	7	33	0	18
0,050	21	0	0	0	20	0	19	18
0,18	7	0	0	0	0	0	0	0
0,18	10	9	7	9	7	7	0	12
0,18	12	21	16	0	3	0	—	0
0,18	14	0	0	17	9	17	—	18
0,18	17	27	24	23	20	23	—	23
0,18	19	34	37	28	21	0	—	33
0,18	21	0	0	0	0	23	—	0
0,64	7	0	0	0	0	0	0	0
0,64	10	5	7	7	7	0	0	6
0,64	12	0	0	21	16	9	8	17
0,64	14	7	11	0	0	17	11	0
0,64	17	28	24	25	20	25	26	17
0,64	19	0	16	32	30	0	0	31
0,64	21	22	0	0	0	30	25	0

Table C.1 (*continued*)

Concentration mg/l	Day	Number of live young produced						
		Rep A	Rep B	Rep C	Rep D	Rep E	Rep F	Rep G
2,3	7	0	0	0	0	0	0	0
2,3	10	0	0	4	0	12	7	0
2,3	12	—	0	0	4	0	0	0
2,3	14	—	11	16	15	19	13	19
2,3	17	—	16	27	21	28	18	27
2,3	19	—	12	0	0	26	0	30
2,3	21	—	0	20	13	0	26	0
8,0	7	0	0	0	0	0	0	0
8,0	10	0	0	0	0	0	0	0
8,0	12	0	0	0	0	0	0	0
8,0	14	0	0	0	0	0	0	0
8,0	17	0	0	—	0	—	—	—
8,0	19	0	—	—	—	—	—	—
8,0	21	0	—	—	—	—	—	—

C.1 Examples of data analysis using hypothesis testing (NOEC determination)

C.1.1 Example 3a — Daphnia chronic reproduction data: Total live young after 14 days exposure

C.1.1.1 NOEC determination by two methods

Since TLY14 is count data, a square-root transform is used. Not reported is an analysis of untransformed values that yielded the same conclusions.

The NOEC exceeds 2,35 mg/l by both methods, the highest concentration for which there was a surviving adult. While it is possible to fit a regression model to these data, given the non-monotone and shallow nature of the dose response, there is little point in doing so.

C.1.1.2 Statistical analysis list file — Ecotoxicity measurements from data set ADAP_REPRO

STATISTICAL ANALYSIS LIST FILE
ECOTOX MEASUREMENTS FROM DATASET ADAP_REPRO
GROUP STATISTICS FOR TLY14 BY DOSE

dose	doseval	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	7	24.0000	21.0	8.4656	3.19970
2	0.015	7	26.2857	24.0	6.0198	2.27527
3	0.053	7	21.4286	22.0	10.6748	4.03471
4	0.19	6	25.3333	25.0	4.2740	1.74483
5	0.67	7	21.2857	23.0	5.4072	2.04374
6	2.35	6	20.0000	19.5	6.3875	2.60768
7	8.23	0	.	.	.	.

C.1.1.3 Shapiro-Wilk test of normality

SHAPIRO-WILK TEST OF NORMALITY OF SQRT(TLY14)

STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
0.75316	-0.43909	0.94111	0.97340	0.45806	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as sole factor. The Shapiro-Wilk test does not indicate a problem with the normality assumption. The Tukey outlier rule is used to identify observations that may be of special interest. Outliers can have an impact on the results, as well as on the assessment of normality and variance homogeneity. A possibly valuable use of these outliers would be to re-run the analysis with outliers omitted to determine whether the NOEC is affected by these outliers. If it is, then caution should be followed in using the results. It is important to understand that just because an observation is identified as an outlier, that does not mean the observation is "bad" or that it should not be used. An observation should be excluded from analysis only for scientifically sound reasons and any such exclusion shall be clearly stated along with its justification.

C.1.1.4 Outliers and influential observations

Outliers and Influential Observations
SQRT(TLY14) FROM FULL DATA

DAPHNIA	dose	doseval	group	OBSER	Pred
20	3	0.053	III	2.23607	4.46961

SE_PRED	L95M	U95M	Resid	LB	UB
0.30488	3.85002	5.08920	-2.23354	-1.61856	1.70612

C.1.1.5 Levene test for SQRT(TLY14) — Full model

LEVENE TEST FOR SQRT(TLY14) - FULL Model

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	5	1.20201	0.32914	

The data was found to be normally distributed (Shapiro-Wilk p-value = 0,458 06) with equal variances (Levene p-value = 0,329 14). An analysis of variance is performed.

Obs	Class	Levels	Values
1	dose	6	1 2 3 4 5 6

C.1.1.6 Overall F-tests for ANOVA

OVERALL F-TESTS FOR ANOVA

Effect	Num DF	Den DF	FValue	ProbF
dose	5	34	0.84	0.5330

The overall ANOVA F-test is not significant. However, this does not affect the remainder of the analysis. As discussed in Clause 5, a significant or non-significant F-test should not be used as a decision rule for the multiple comparisons of step-down Jonckheere-Terpstra test. The F-test can be affected by significant differences among treatments of no interest to toxicology or by the necessity to control for a large number of possible differences of no interest to toxicology.

C.1.1.7 Estimated dose effects and Dunnett for decreasing Alternative using Alpha = 0,05

```
ESTIMATED DOSE EFFECTS and DUNNETT FOR Decreasing ALTERNATIVE
        USING ALPHA=.05 FOR COMPARISONS TO CONTROL
```

		Dunnett one-sided *p*-value	Test Group Mean	N
Estimate	SIGNIF			
DOSE TREND		0.24959	.	.
DOSE QUAD		0.65360	.	.
DOSE 2-1		0.95623	5.09841	7
DOSE 3-1		0.49052	4.46961	7
DOSE 4-1		0.92891	5.01786	6
DOSE 5-1		0.60723	4.57826	7
DOSE 6-1		0.45937	4.42441	6

The Dose Trend and Dose Quad are formal tests for departure from monotonicity in the dose response. They are based on linear contrasts, as described in 5.1.4. In this instance, the test for linear dose response is not significant but neither is the test for departure from linearity (Dose Quad), so there is no reason, based on these tests, not to go on with the step-down Jonckheere-Terpstra test. However, an inspection of the treatment means does indicate some non-monotonicity in the dose response. Thus, some caution should be used in interpreting the results of the Jonckheere-Terpstra test presented below.

In the results presented below,

— JONC is the value of the Jonckheere-Terpstra test statistics;

— P1DNCF is the *p*-value associated with this test statistic to test the hypothesis of a downward trend;

— P1UPCF is the test statistic for testing the significance of an upward trend and is a default printout of the software used and is not utilized in the present analysis, i.e. the *p*-value for the upward trend;

— ZC is a standardized value of the JONC statistic, computed with tie correction;

— ZCCF is ZC, except that it is computed using a standard continuity correction factor.

The CF in several terms refers to the use of this continuity correction factor.

p-values are for the tie-corrected test with the continuity correction factor.

SIGNIF RESULTS are for a decreasing alternative hypothesis.

```
Jonckheere Trend Test on Dose 0 + Lowest 5 Doses through 2.35 mg/l
Analysis of TLY14
```

JONC	ZC	ZCCF	P1UPCF	P1DNCF	SIGNIF
281.5	-1.223426	-1.211548	0.8871572	0.1083588	

Since the Jonckheere-Terpstra test with all concentrations present is not significant, no further testing is required. Jonckheere test results are included in the summary table. Group means should be examined to check for lack-of-fit to a linear trend before trend test results are accepted.

It is observed that neither Dunnett's test nor the Jonckheere-Terpstra test found a significant effect at any concentration at 14 days.

C.1.2 Example 3b — Daphnia chronic reproduction data: Total live young (TLY) after 21 days exposure

C.1.2.1 NOEC determination by two methods

Since TLY21 is count data, a square-root transform is used. Not reported is an analysis of untransformed values that yielded the same conclusions.

The NOEC by Dunnett's test exceeds 2,35 mg/l, the highest tested concentration for which at least one adult daphnia survived. The NOEC by the Jonckheere-Terpstra test is 0,19 mg/l.

C.1.2.2 Ecotoxicity measurements from data set ADAP_REPRO — Group statistics for TLY21 by dose

ECOTOX MEASUREMENTS FROM DATASET ADAP_REPRO
GROUP STATISTICS FOR TLY21 BY DOSE

dose	doseval	COUNT	MEAN	MEDIAN	STD_DEV	STD_ERR
1	0	7	84.5714	87.0	20.3130	7.67760
2	0.015	7	86.5714	89.0	11.8583	4.48201
3	0.053	7	72.2857	84.0	21.2581	8.03479
4	0.19	6	78.0000	80.5	11.4717	4.68330
5	0.67	7	71.4286	71.0	9.5718	3.61779
6	2.35	6	64.0000	65.5	16.3707	6.68331
7	8.23	0	.	.	.	.

C.1.2.3 Shapiro-Wilk test of normality

SHAPIRO-WILK TEST OF NORMALITY OF SQRT(TLY21)
AQUATIC DAPHNIA: FULL DATA

STD	SKEW	KURT	SW_STAT	P_VALUE	SIGNIF
0.87309	-0.34552	-0.55535	0.96350	0.22028	

The Shapiro-Wilk test was done on the residuals from a simple 1-factor ANOVA with concentration as the sole factor. The Shapiro-Wilk test does not indicate a problem with the normality assumption. The Tukey outlier rule was used to identify observations that may be of special interest. For these data, no outliers were identified.

C.1.2.4 Levene test

LEVENE TEST FOR SQRT(TLY21) - FULL Model
ANALYSIS OF VARIANCE ON FULL DATA SET

Effect	DF	LEVENE	P_VALUE	SIGNIF
DOSE	5	0.91555	0.48269	

The data was found to be consistent with a normal distribution with equal variances. An analysis of variance is performed.

Obs	Class	Levels	Values
1	dose	6	1 2 3 4 5 6

C.1.2.5 Overall F-tests for ANOVA

OVERALL F-TESTS FOR ANOVA

Effect	Num DF	Den DF	FValue	ProbF
dose	5	34	1.90	0.1199

The overall ANOVA F-test is not significant. However, this does not affect the remainder of the analysis. As discussed in Clause 5, a significant or non-significant F-test should not be used as a decision rule for the multiple comparisons of step-down Jonckheere-Terpstra test. The F-test can be affected by significant differences among treatments of no interest to toxicology or by the necessity to control for a large number of possible differences of no interest to toxicology.

C.1.2.6 Estimated dose effects and Dunnett for decreasing alternative using alpha = 0,05

ESTIMATED DOSE EFFECTS and DUNNETT FOR Decreasing ALTERNATIVE
USING ALPHA =.05 FOR COMPARISONS TO CONTROL

Estimate	SIGNIF	Dunnett one-sided *p*-value	Test Group Mean	N
DOSE TREND	*	0.01150	.	.
DOSE QUAD		0.71556	.	.
DOSE 2-1		0.91014	9.28556	7
DOSE 3-1		0.24321	8.41734	7
DOSE 4-1		0.59511	8.81095	6
DOSE 5-1		0.25588	8.43515	7
DOSE 6-1		0.05306	7.94130	6

The Dose Trend and Dose Quad are formal tests for departure from monotonicity in the dose response. They are based on linear contrasts, as described in 5.1.4. In this instance, the test for linear dose response is significant but not the test for departure from linearity (Dose Quad), so there is no reason, based on these tests, not to go on with the step-down Jonckheere-Terpstra test. As discussed in 5.3, the only condition, based on these formal tests, to question the use of the Jonckheere-Terpstra test would be a non-significant test for dose trend and a significant test for dose quad. In addition, an inspection of the treatment means indicate some non-monotonicity in the dose response, but the overall downward trend is quite evident.

In the results presented below,

— JONC is the value of the Jonckheere-Terpstra test statistics;

— P1DNCF is the *p*-value associated with this test statistic to test the hypothesis of a downward trend;

— P1UPCF is the test statistic for testing the significance of an upward trend and is a default printout of the software used and is not utilized in the present analysis, i.e. the *p*-value for the upward trend;

— ZC is a standardized value of the JONC statistic, computed with tie correction;

— ZCCF is ZC, except that it is computed using a standard continuity correction factor.

The CF in several terms refers to the use of this continuity correction factor.

p-values are for the tie-corrected test with the continuity correction factor.

SIGNIF RESULTS are for a decreasing alternative hypothesis.

```
Jonckheere Trend Test on Dose 0 + Lowest 5 Doses through 2.35 mg/l

             JONC        ZC       ZCCF     P1UPCF     P1DNCF SIGNIF
            216.5 -2.761099 -2.749249 0.9970134 0.0027775 **

Jonckheere Trend Test on Dose 0 + Lowest 4 Doses through 0.67 mg/l

             JONC        ZC       ZCCF     P1UPCF     P1DNCF SIGNIF
            163.5 -2.050218 -2.035031 0.9790761 0.0194424  *

Jonckheere Trend Test on Dose 0 + Lowest 3 Doses through 0.19 mg/l

             JONC        ZC       ZCCF     P1UPCF     P1DNCF SIGNIF
            107.5 -1.255247 -1.233604 0.8913248 0.1008208
```

Since the Jonckheere-Terpstra test is not significant with all concentrations above 0,19 mg/l omitted, no further testing is required and the NOEC is 0,19. It is observed that Dunnett's test found no significance at any concentration.

The Jonckheere test results are included in the summary table. Group means should be examined to check for lack-of-fit to a linear trend before trend test results are accepted.

C.2 Example of data analysis by dose-response modelling

C.2.1 Daphnia reproduction test: number of young

Total number of life young are reported at various points in time after exposure, showing a clear increase in rate of number of young produced with time. This information in time may be used, as shown at the end of this subclause. First, however, it is illustrated how the number of young for a particular time period may be analysed by dose-response modelling

Since the production of young only starts after a number of days, with an increasing rate in the period thereafter, the total count over the first two weeks, or that over the third week may, for example, be chosen as the response variable for dose-response analysis. As Figure C.1 illustrates, the production of young has indeed increased with age. This figure also illustrates that the dose-response relationship of the counts over the first two weeks is similar to that over the third week. Although in this way the data from the first two weeks and from the third week are used together in a single analysis, the problem is that the analysis assumes the data to be independent, which they are not (see Figure C.2). Therefore, the analysis of Figure C.1 is from a statistical point of view not valid, in particular the confidence interval may not be reliable.

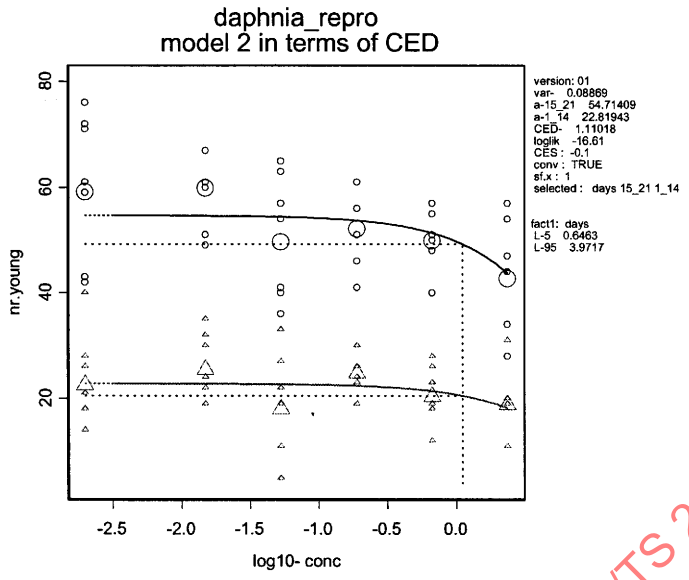


Figure C.1 — Number of life young as a function of concentration (on log-scale to improve visibility), counted over the first two weeks (triangles) and over the third week (circles)

CED = EC_{10}. Model 2: $y = a \exp(bx)$. This model was reparameterized by substituting the parameter b by the EC_{10}. In this way, the confidence interval for the EC_{10} (L-5 to L-95) can be estimated by the likelihood profile method. Note that the confidence interval in this analysis may not be reliable, due to violation of independence of the data (see Figure C.2).

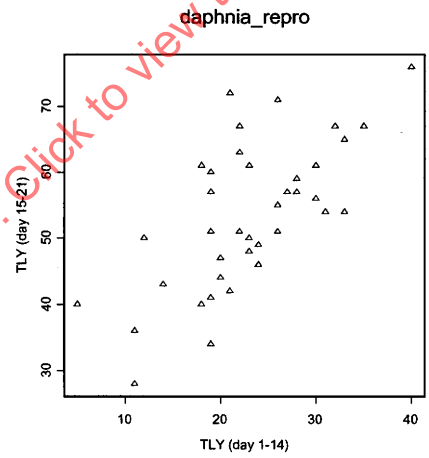


Figure C.2 — Total live young (TLY) in third week plotted against TLY in first two weeks, showing the correlations between these counts

The obvious way to avoid the problem of dependent data in the analysis of Figure C.1 is to perform the analysis on the counts over the third week only. Another argument for this selection is that it may be assumed that at this time the reproduction rate has reached a more or less stable level, whereas the counts over the first two-weeks period includes the starting-up of the reproduction (leading to more variation and problems of interpretation). According to the recommendation of Clause 6, various models are fitted to the counts over week three.

First the nested non-linear model proposed by Slob (2002) is fitted. This results in the following log-likelihoods:

Model 1	$y = a$	loglik $= 4{,}19$
Model 2	$y = a \exp(x/b)$	loglik $= 8{,}26$
Model 3	$y = a \exp(\pm(x/b)^d)$	loglik $= 8{,}26$
Model 4	$y = a\,[c - (c - 1) \exp(-x/b)]$	loglik $= 8{,}26$

The log-likelihoods can be compared with the likelihood-ratio test. A model with one more parameter fits the data significantly better (at $\alpha = 0{,}05$) than the model without that parameter when the increase in the log-likelihood is greater than 1,92. The difference in log-likelihoods between Model 2 and Model 1 is 4,07, so it may be concluded that the data show a significant dose response. The models with more parameters result in the same log-likelihood, and it may be concluded that Model 2 is, from this family of models, the appropriate one for describing the data. Figure C.3 shows the results for this model.

Next a polynomial model is fitted to these data. Since this is again a nested family of models, the log-likelihoods can be compared by the ratio-likelihood test:

Model 1	$y = a$	loglik $= 4{,}19$
Model 2	$y = a + bx$	loglik $= 8{,}20$
Model 3	$y = a + bx + cx^2$	loglik $= 8{,}50$

Here, Model 2 (straight line) is not significantly improved by higher-order polynomials, and this model may be selected from this family of models.

Finally, the power model, $y = c + ax^b$, is fitted to the number of young. This model results in a log-likelihood value of 9,15. However, by fixing the parameter b to one, the model reduces to a straight line. Hence, the power model and the straight line are nested, and their log-likelihoods can be compared. Since the straight line resulted in a log-likelihood of 8,20, the power model does not give a significantly better fit.

It may be concluded that these data should be described by a two-parameter model, either the exponential, or the linear (straight-line) model. Figure C.3 shows the fit of the exponential model. Table C.2 summarizes the results for both fits, showing that the EC_{10} (and its confidence interval) are similar for both models.

Table C.2 — Number of live young: Summary of results for exponential and straight-line model, both having two parameters

Model	Log-lik	EC_{10}	90 % CI
$y = a \exp(bx)$	8,26	0,91	0,58 to 2,05[a]
$y = a + bx$	8,20	1,00	0,69 to 2,15[b]

[a] Based on profile-likelihood method.

[b] Based on 1 000 bootstrap runs.

C.2.2 Assumptions

To check the assumptions of normality and homogeneous variances, the regression residuals (i.e. the deviations of the individual data points from the dose-response model) may be plotted in various ways. Here, two plots are considered. One is the so-called QQ-plot, where the observed quantiles are plotted against the theoretical quantiles, e.g. according to the normal distribution. When data are sampled from a normal distribution, this plot should theoretically result in a straight line. It should be noticed that fitting a line to a QQ-plot is unsound (which is not always recognized). One may draw the theoretical straight line in the plot, with

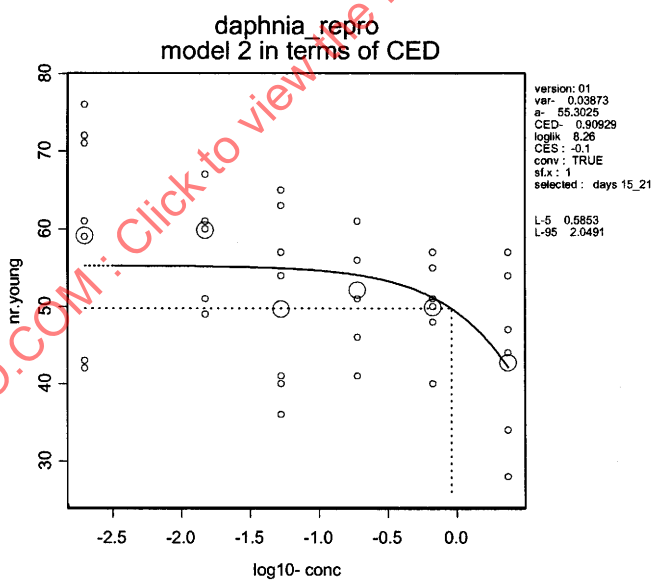
intercept equal to the mean of the data points and with slope equal to the standard deviation of the data points. In the case of regression, the data points are the regression residuals, which are corrected for the dose-response relationship.

In interpreting a QQ-plot, one should realize that, due to sampling errors, fluctuations around the line can easily arise, especially in small data sets. In particular, a pattern resembling Aesculapius' staff is not unusual, even for data that are sampled from a normal distribution by the computer. Hence, QQ-plots should only lead to the conclusion that the assumed distribution is inadequate when the data show a clear overall curvature. It is always the general trend, not single data points that should be considered.

As Figure C.4 (upper left panel) shows, the data did comply with the assumption of log-normality, since these residuals resulted from an analysis on the log-counts. The same residuals plotted against concentration (upper right panel) do not reveal a clear trend, and the assumption of homogeneous variances (on log-scale) appears acceptable.

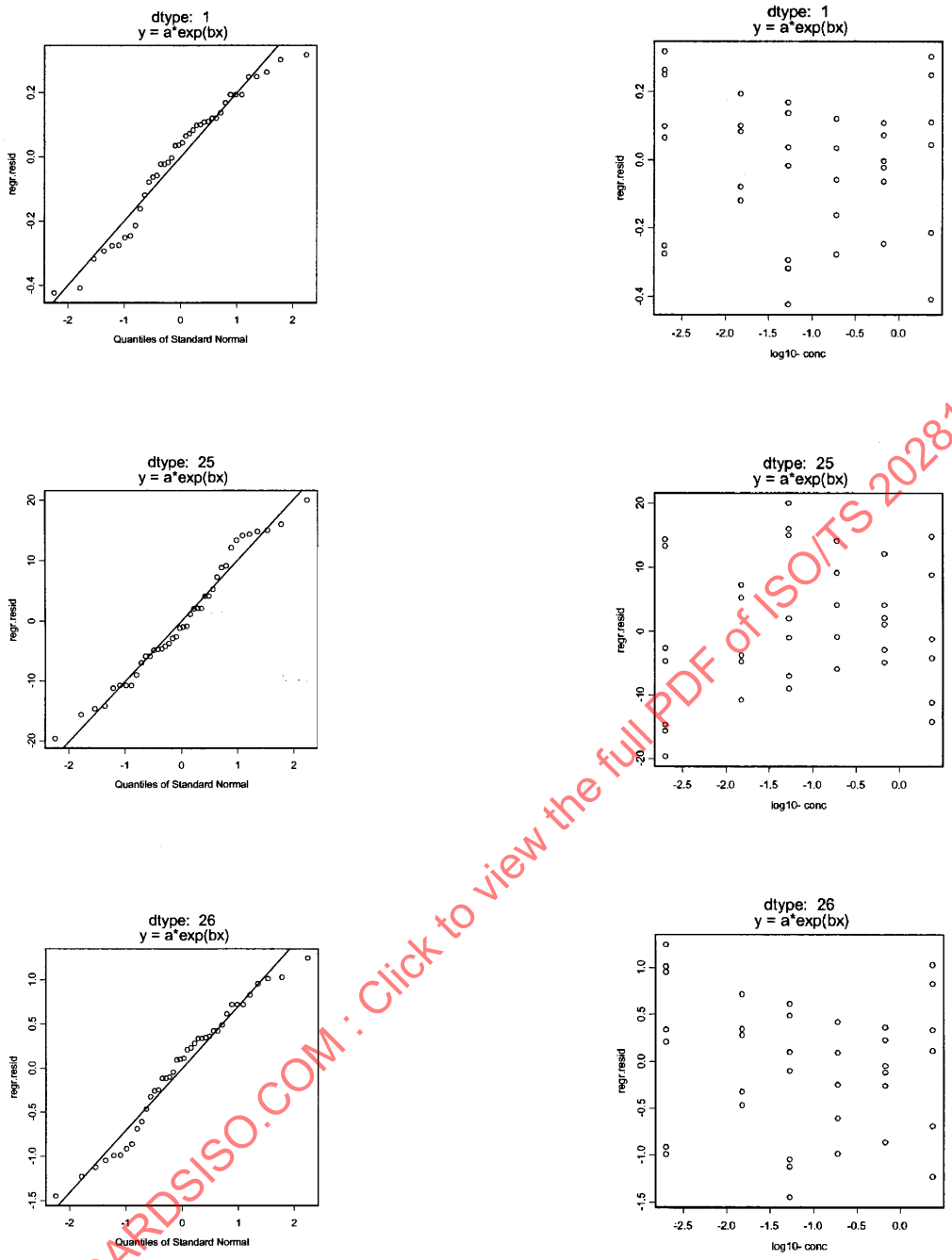
Although not strictly needed, the residual plots are also shown for an analysis where the log-transformation was omitted (middle panels of Figure C.4), as well as where a square root transformation was applied (lower panels of Figure C.4). The plots for these three situations are similar. The reason of this similarity is that the scatter in these data is relatively small (CV of around 20 %) A log-normal distribution gets closer to a normal distribution with smaller variation (CV).

Therefore, the smaller the scatter in the data, the more data are needed to see any difference in the QQ-plots assuming normality, log-normality, or square-root-normality. For the same reason, it may be expected that applying or omitting any transformation has no large impact on the results of the analysis when the scatter in the data is relatively small. Indeed, re-analysing these data without transformation results in an EC_{10} = 0,90 mg/l, while the same analysis with log-transformation resulted in EC_{10} = 0,91 mg/l.



NOTE This model was reparameterized by substituting the parameter b by the EC_{10}. In this way, the confidence interval for the EC_{10} (L – 5 to L – 95) can be estimated by the likelihood profile method.

Figure C.3 — Exponential model fitted to the number of life young counted over week three



a) QQ-plots for regression residuals for the exponential model

b) The same residuals plotted against dose (dose on log-scale to improve visibility)

NOTE Upper panels: analysis on log-scale.

Middle panels: no transformation.

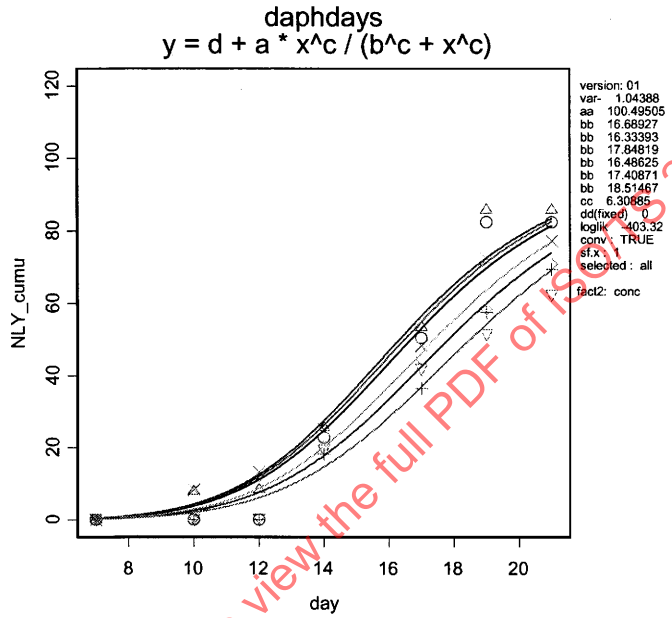
Lower panels: analysis on square root scale.

Figure C.4 — Plots of regression residuals

C.2.3 Two-step analysis: Taking time into account

Similar to the algal test example, these data can be analysed in two steps, taking the information in time into account. In the first step, the (cumulative) number of eggs is considered as a function of time. Figure C.5 shows the Hill model fitted to the data, where for each concentration a separate value is estimated for the parameter b (= ET_{50}, time at which 50 % of maximum value is achieved).

In these data of cumulative counts, the variance clearly increases with the mean (data not shown). A log-transformation resulted in the variance decreasing with the mean, but a square root transformation resulted in homogeneous variance, and compliance with the normal distribution.



NOTE Each symbol represents a concentration, and to each concentration the Hill model is fitted, assuming that the parameters a and c are equal amongst concentrations, while parameter b was assumed to be different. The parameter d was set at zero here (since reproduction is zero at time zero).

Figure C.5 — Means of number of young, plotted cumulatively against time

In the second step, the estimated values for the ET_{50} are considered as a function of the concentration, and a dose-response model is fitted to these data (see Figure C.6). Again, the second (nonzero) concentration appears to deviate from the general pattern (compare with Figure C.3). The EC_{10} for this endpoint is estimated at 2,30 mg/l, with a confidence interval (1,41 to 6,19).

Since the number of young are followed in time for a number of replicates at each concentration, it is better to estimate an ET_{50} for each replicate (see discussion in algal data set). This is illustrated in Figure C.7. Next the ET_{50} values for each replicate may be considered as a function of the concentration (see Figure C.8). The EC_{10} is estimated at 2,36 mg/l, similar to the value obtained in Figure C.6, but the confidence interval is somewhat larger (1,36 to 8,82).

From Figure C.8 it becomes apparent that the deviation of the second concentration group is caused by two outlier replicates. When these two outliers are removed, a more regular dose-response relationship results (see Figure C.9). This also results in a lower EC_{10} (1,90 mg/l), and a smaller confidence interval (1,30 to 3,49).

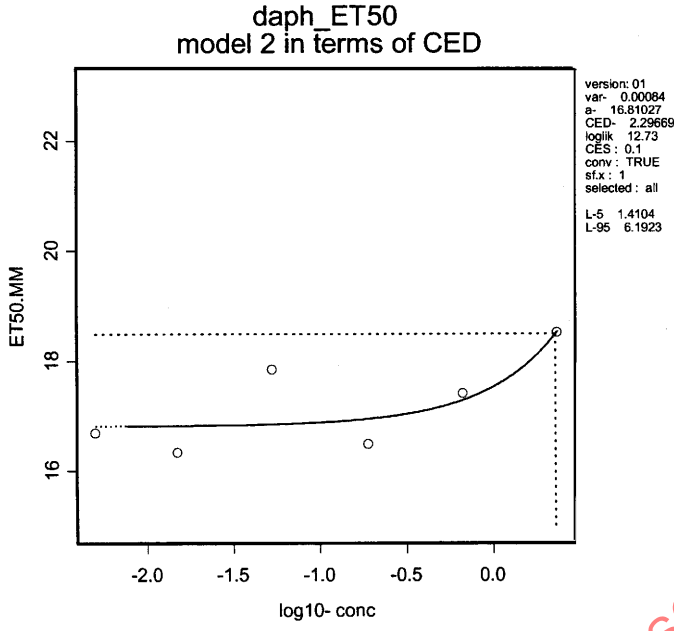
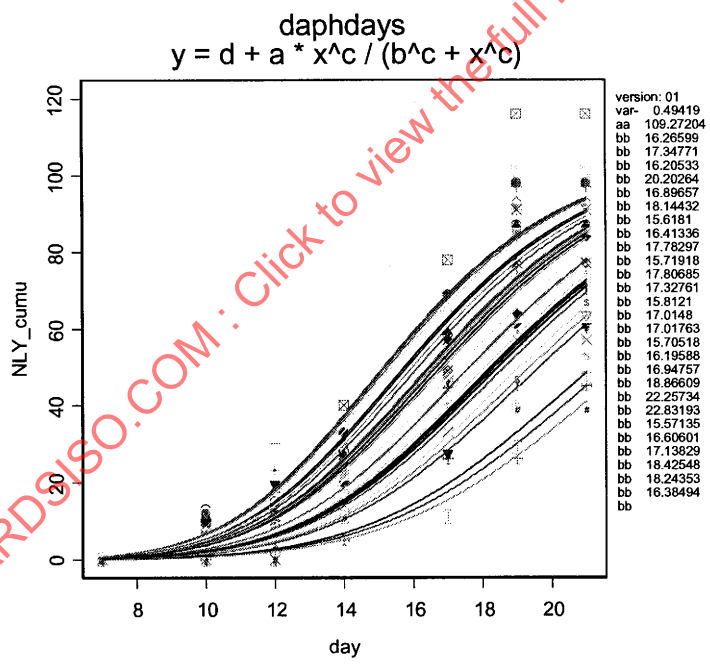


Figure C.6 — Estimated ET_{50} values from Figure C.5, plotted against the concentration, with a fitted dose-response model



NOTE Each symbol represents a replicate, and to each replicate the Hill model is fitted, assuming that the parameters a and c are equal amongst replicates, while parameter b was assumed to be different. The parameter d was set at zero here (since reproduction is zero at time zero).

Figure C.7 — Number of young, plotted cumulatively against time

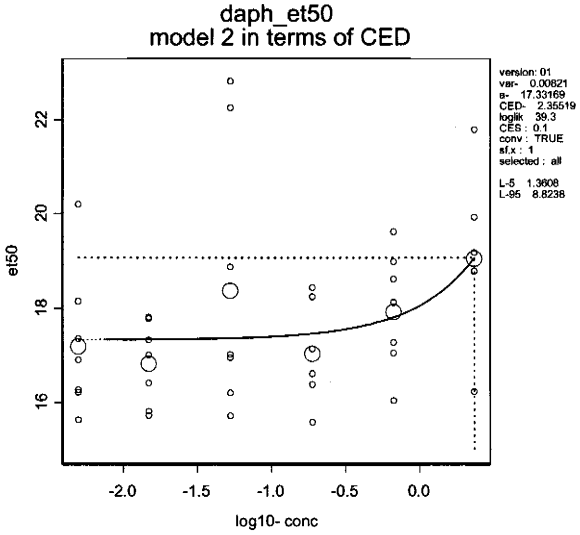


Figure C.8 — ET_{50}s estimated per replicate (see Figure C.7) as a function of the concentration with a fitted dose-response model

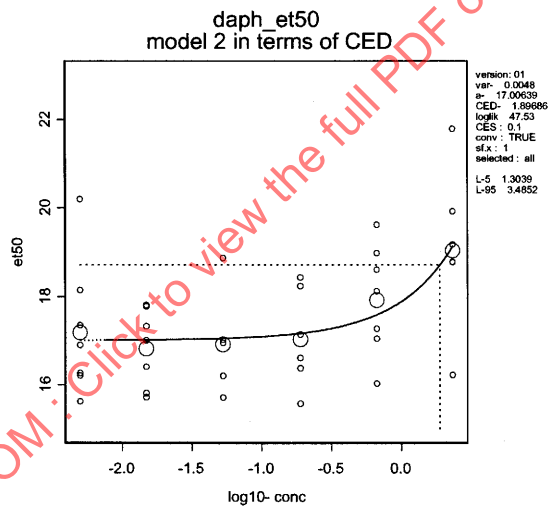


Figure C.9 — ET_{50}s estimated per replicate (see Figure C.7) as a function of the concentration, with two outliers removed

C.3 Examples of data analysis using DEBtox (biological methods)

C.3.1 Data

Figure C.10 shows data for the cumulative number of offspring per female as affected by an unknown compound. The data were weighted in the estimation of parameters by the number of surviving females (data not given here).

	Time: day, Conc: milligram/liter, Resp: Number of offspring						
	0.000	0.015	0.053	0.190	0.670	2.350	8.230
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
7	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
10	7.14286	8.28571	6.28571	7.28571	4.57143	3.28571	0.00000
12	9.42857	8.85714	9.14286	13.95240	14.71430	3.95238	0.00000
14	24.00000	26.28570	21.42860	24.11900	21.28570	19.45240	0.00000
17	52.00000	53.85710	42.42860	47.45240	44.85710	42.28570	0.00000
19	84.57140	86.57140	64.14290	72.95240	60.42860	53.61900	0.00000
21	84.57140	86.57140	72.28570	76.78570	71.42860	63.45240	0.00000

Figure C.10 — DEBtox example: Data for the cumulative number of offspring per female as affected by an unknown compound

C.3.2 Parameter estimates and asymptotic standard deviations (ASD)

Reproduction, Maintenance model			ASD	Correlation coefficients		
No effect concentration	3.895e-009	mg l^{-1}	0.004			
Tolerance concentration	0.2265	mg l^{-1}	35.175	0.233		
Maximal reproduction rate	15.9	No d^{-1}	0.646	-0.672	-0.030	
Elimination rate	0.001268	d^{-1}	0.199	0.233	1.000	-0.031
Von Bertalanffy growth rate	0.1	d^{-1}				
Scaled length at birth	0.13					
Scaled length at puberty	0.61					
Energy investment ratio	1					
Mean deviation	5.207					

Figure C.11 — DEBtox example: Parameter estimates and asymptotic standard deviations (ASD)

Table C.3 — EC_x values (derived from parameter values)

Values in milligrams per litre

Day	EC_0	ASD	EC_{50}	ASD
21	$1{,}10^{-5}$	0,44	4,22	6,59

C.3.3 Graphical test of model predictions against data

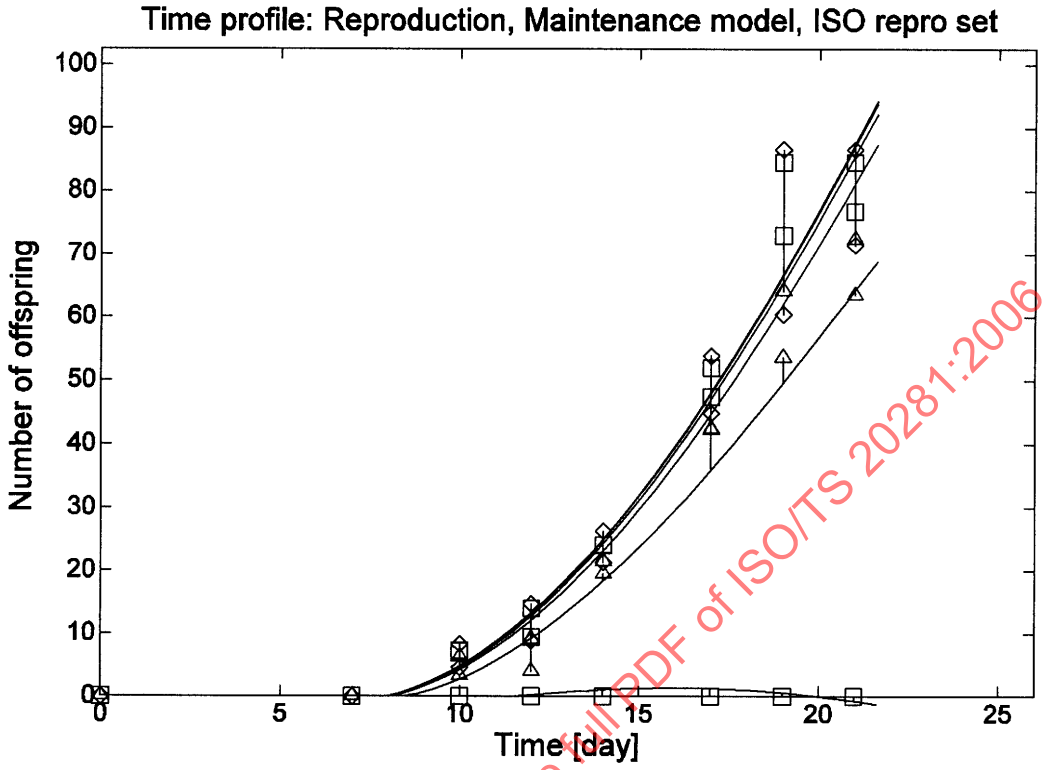


Figure C.12 — DEBtox example: Time profile (reproduction, maintenance model, ISO repro set)

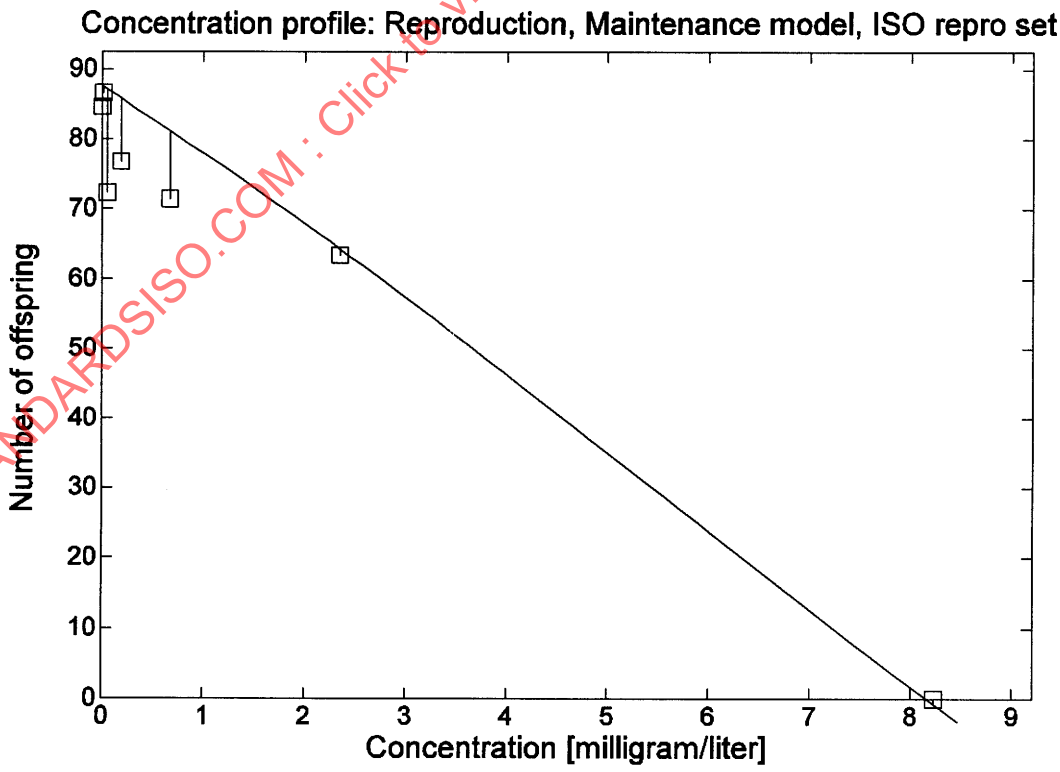


Figure C.13 — DEBtox example: Concentration profile (reproduction, maintenance model, ISO repro set)

C.3.4 Profile likelihood for NEC estimate

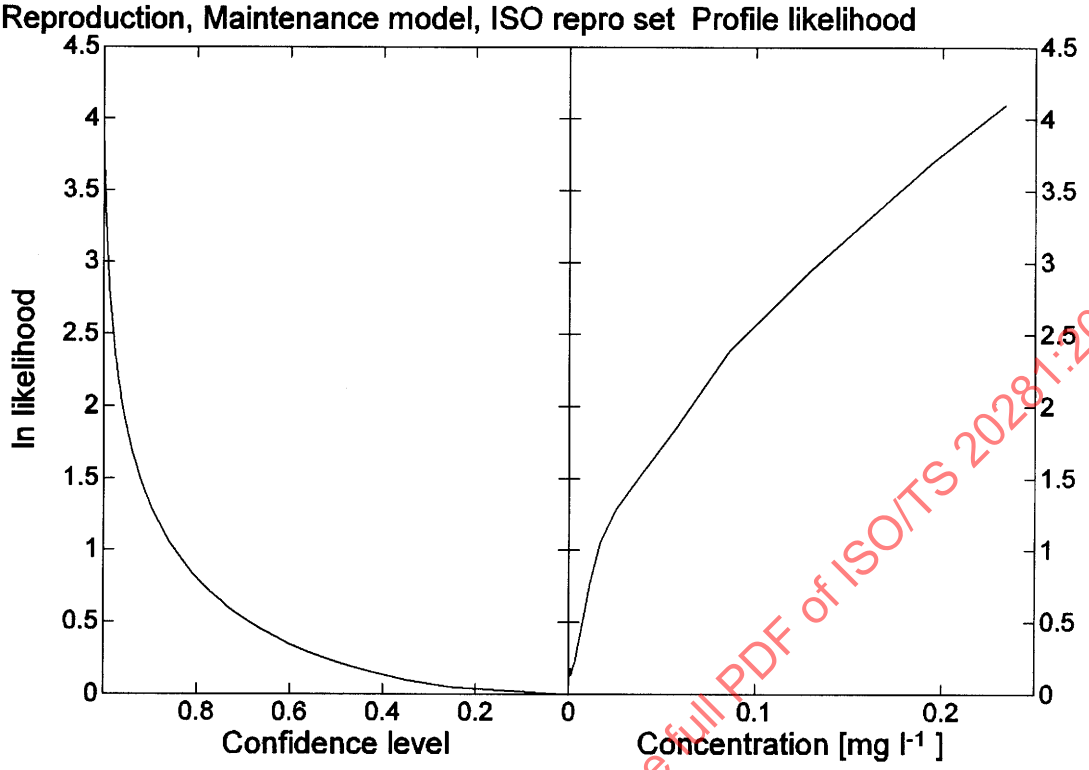


Figure C.14 — DEBtox example: Profile likelihood for NEC estimate (reproduction, maintenance model, ISO repro set)

C.3.5 Body length at 21 days

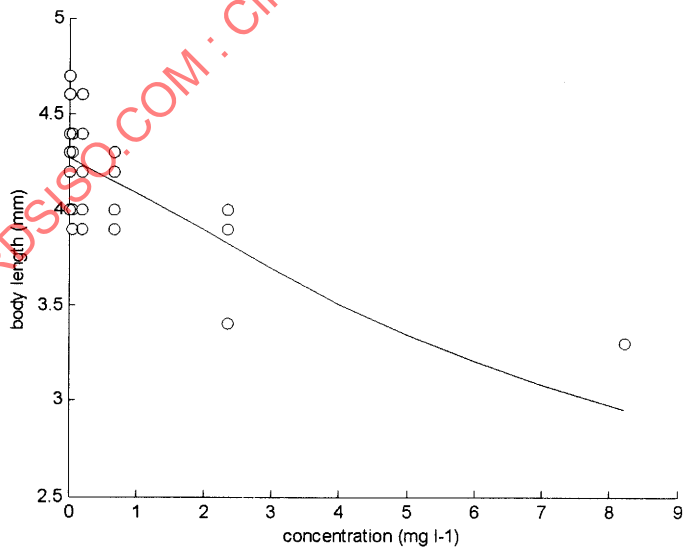


Figure C.15 — DEBtox example: Body length at 21 days

C.3.6 Comments

This dataset is special in several respects. We have counts of offspring at relatively few points in time, not for each day as the guideline recommends. This reduces the effectiveness of the biology-based method; the fact that these data do not give detailed information about the start of the reproduction is especially troublesome.

The first graph (Figure C.12) shows that reproduction starts here later than expected on the basis of the default value of 0,42 for the scaled length at puberty. Therefore, this scaled length has been set at 0,61 to mimic this late start. The model for effects on maintenance appeared to fit the data best; an increase in maintenance costs reduces the ultimate length of the daphnia.

This is confirmed by the length data, shown in the last plot (Figure C.15); the fitted length at 21 d are calculated with DEBtool; the plotted curve involves the estimation of a single parameter: the ultimate length in the blank. All other parameters are already fixed by the reproduction data, and determine the toxico-kinetics, including the dilution by growth, and the effects on growth during exposure. The length data were not used to estimate any effect parameters. The good fit of the length data confirms the effects on growth by an increase of the maintenance costs as expected from the observed effects on reproduction. Direct effects on reproduction would not affect body size; direct effects on growth would affect the growth rate, but not the ultimate size. The data do not give information about growth, but it is likely that growth almost ceased before 21 d for daphnia, even in the stressed situation. The NEC was found to be not significantly different from zero, with a 95 % confidence interval of (0 to 0,082) mg/l.

Annex D
(informative)

Analysis of a “fish growth” data set (OECD GL 204/215 – ISO 10229) using the three presented approaches

D.1 Data set

The data used here have been obtained for a 21-day fish test following OECD GL204.

Test organism: *Oncorhynchus mykiss*

Temperature: 14 °C to 15 °C

Table D.1 — Number of dead fish

	Control	Nominal concentration of the test item mg/l					
Day	0	1,0	2,2	4,6	10	22	46
1							
2							
3							
6							1
7							2
8							2
9							2
10			1				5
13			1			1	5
14			1		1	1	5
15			1		1	1	5
16			1		1	1	5
17			1		1	1	5
21	0	0	1	0	1	1	6

Table D.2 — Length of the fish in millimetres

	Control	Nominal concentration of the test item mg/l					
	0	1,0	2,2	4,6	10	22	46
0 days	5,7	5,7	5,1	5,9	5,9	6,0	5,9
	5,3	5,8	6,0	5,9	5,7	5,6	5,6
	6,0	5,6	5,7	5,7	5,5	5,9	6,0
	5,9	6,0	5,8	5,5	5,6	5,6	5,6
	5,9	5,5	5,5	6,0	5,6	5,7	5,3
	5,6	6,0	5,8	6,0	5,9	5,8	6,0
	5,6	5,9	6,0	5,6	6,0	5,2	5,0
	5,0	5,1	5,6	5,3	5,0	5,7	5,3
	6,0	5,2	5,1	5,3	5,5	5,1	5,1
	5,4	5,4	5,6	5,1	5,1	5,6	5,5
28 days	6,5	6,2	6,5	6,5	6,7	6,3	5,3
	6,5	6,9	6,4	6,5	6,3	6,0	6,3
	6,4	6,8	6,5	6,9	5,8	6,9	5,2
	7,1	6,0	6,6	6,2	5,7	5,7	4,9
	6,8	5,8	6,5	6,0	6,4	6,7	*
	5,4	6,7	7,3	6,4	5,7	6,7	*
	6,6	6,9	6,8	5,8	6,2	5,9	*
	6,3	6,3	6,1	7,2	6,8	6,2	*
	6,1	6,3	6,0	6,0	6,1	5,8	*
	6,7	6,5	*	6,0	*	*	*

Table D.3 — Wet mass in grams

	Control	Nominal concentration of the test item mg/l					
	0	1,0	2,2	4,6	10	22	46
0 days	1,9	1,8	1,4	1,8	2,2	2,2	2,2
	1,7	2,0	2,2	2,0	1,9	1,7	1,9
	2,7	1,9	1,8	2,1	1,8	2,1	2,3
	2,0	2,2	2,1	1,8	2,2	1,8	1,8
	1,8	1,6	1,9	2,0	1,7	1,9	1,7
	1,8	2,2	1,9	2,3	2,0	2,2	2,0
	1,9	1,8	2,4	1,9	2,1	1,5	1,4
	1,4	1,6	1,7	1,7	1,4	1,8	1,5
	2,3	1,5	1,5	1,5	1,5	1,5	1,6
	1,5	1,5	1,7	1,5	1,4	1,7	1,6
28 days	2,8	2,7	2,9	2,9	3,4	2,7	1,6
	3,0	3,3	2,6	3,0	2,8	2,0	2,8
	2,7	2,9	2,7	3,5	2,1	3,5	1,2
	3,9	2,2	3,3	2,7	2,3	1,8	0,9
	3,1	2,0	2,7	2,3	3,1	3,1	*
	1,8	3,1	4,0	2,7	1,8	3,2	*
	2,9	3,2	3,0	2,0	2,4	2,2	*
	2,5	2,5	2,5	4,0	3,0	2,5	*
	2,2	2,5	2,2	2,2	2,3	1,8	*
	3,1	2,6	*	2,4	*	*	*

D.2 Examples of data analysis using hypothesis testing (NOEC determination)

D.2.1 NOEC determination for growth

Given that there are two measurements on each fish, one on Day 0 (before exposure to compound) and one on Day 28, this is a repeated-measures experiment. It is unfortunate that individual fish are not identified. This means there is no way to determine growth on an individual fish and statistical power may suffer as a result. No replicate information is provided, so it is not possible to treat the replicate as sampling unit and do repeated measures analysis or paired difference analysis on the replicate means. The data cannot be analysed therefore with repeated measures methodology. There are nonetheless at least two ways to proceed to determine the NOEC.

— **Method 1**. Compute treatment means for each day and concentration. The response to be analysed is the difference of the treatment means for each concentration, to obtain the mean growth from day 0 to day 28 for that treatment group. This leaves one observation per treatment group. ANOVA methods, such as Dunnett's test, are then not available, since there is no estimate of error. However, the Jonckheere-Terpstra test can still be applied, preferably in its exact permutation implementation. This approach ignores the reduced sample size in the 28-day high concentration group.

— **Method 2**. Do a two-factor ANOVA with day and concentration, and their interaction as the model terms. Then compare the growth, (day 28 mean) – (day 0 mean), in each concentration to the growth in the control by standard ANOVA methods. This is not entirely correct, since it ignores the repeated measures nature of the data.

In the present case, the two lead to the same NOEC, namely the 22 mg/l concentration. Of the two, method 1 is theoretically soundest. Details are provided below.

D.2.2 Length

D.2.2.1 Method 1

The step-down Jonckheere-Terpstra test is applied, first with all concentrations present, then with the 46 mg/l group omitted. The variable DeltaL is the difference of the day 28 mean minus the day 0 mean for each concentration. This was done using SAS Proc Freq. The default output includes both one- and two-sided tests for trend and both exact and asymptotic tests. These are all left for the reader to see, but only the one-sided exact results are used in the discussion.

```
Jonckheere-Terpstra Test of DeltaL Through Conc=46

      Statistics for Table of DeltaL by Conc

            Jonckheere-Terpstra Test
      ------------------------------------
      Statistic (JT)                 3.0000
      Z                             -2.1268

      Asymptotic Test
      One-sided Pr <  Z              0.0167
      Two-sided Pr > |Z|             0.0334

      Exact Test
      One-sided Pr <=  JT            0.0218
      Two-sided Pr >= |JT - Mean|    0.0437
```

Since the Jonckheere-Terpstra test with all concentrations included is significant at the 0,05 level (p-value for exact one-sided trend is = 0,0218), the high concentration is omitted and the test is repeated with the remaining concentrations.

Jonckheere-Terpstra Test of DeltaL Through Conc=22

Statistics for Table of DeltaL by Conc

Jonckheere-Terpstra Test	
Statistic (JT)	3.0000
Z	-1.5302
Asymptotic Test	
One-sided Pr < Z	0.0630
Two-sided Pr > |Z|	0.1260
Exact Test	
One-sided Pr <= JT	0.0944
Two-sided Pr >= |JT - Mean|	0.1889

This test is not significant at the 0,05 level, so no further testing is required and the NOEC is 22 mg/l.

D.2.2.2 Method 2

D.2.2.2.1 Full trout-size data set and basic statistics for length

Fish Growth Example: Length
Trout Size Data
FULL DATA SET

Method 2 is a two-factor ANOVA on length, with day and concentration, and their interaction as model terms. First, the simple means are computed. It is apparent that there is a dose response, perhaps beginning with the 2,2 mg/l concentration.

Obs	Conc	DeltaL	DeltaW
1	0	0.94000	1.20000
2	1	0.94000	1.10000
3	2.2	1.02222	1.27778
4	4.6	0.85000	1.17000
5	10	0.68889	0.97778
6	22	0.74444	0.93333
7	46	-0.07500	0.02500

Basic Statistics for length

Days	Conc	mean_ length	std_length	n_length
0	0	5.64	0.33065591	10
0	1	5.62	0.3190263	10
0	2.2	5.62	0.3190263	10
0	4.6	5.63	0.32335052	10
0	10	5.58	0.32930904	10
0	22	5.61	0.28460499	10
0	46	5.53	0.3591657	10
28	0	6.44	0.4575296	10
28	1	6.44	0.38355066	10
28	2.2	6.52	0.38005848	9
28	4.6	6.35	0.44284434	10

28	10	6.19	0.40756731	9
28	22	6.24	0.43620841	9
28	46	5.425	0.60759087	4

It is observed that mortality in the high concentration has noticeably affected the sample size on day 28. The other differences in sample size are quite small and are likely to have insignificant impact on conclusions.

CovParm	Estimate
Residual	0.1428

Effect	Num DF	Den DF	FValue	ProbF	MSERR	SSQRS	SSERR
Days	1	117	84.86	<.0001	0.14282	12.1195	16.7102
Conc	6	117	3.78	0.0018	0.14282	3.2401	16.7102
Days*Conc	6	117	2.57	0.0224	0.14282	2.2045	16.7102

These are the overall F-tests for ANOVA. It is observed that both main effects and the interaction are significant at the 0,05 level. The CovParm above the F-tests is the pooled sample mean-squared error.

D.2.2.2.2 ANOVA summary statistics

ANOVA SUMMARY STATISTCS

MODELSS	SSERR	TOTSS	RSQUARE
20.1477	16.7102	36.8579	0.54663

These are basic measures that can be used for model assessment. Since this is a linear model, the R-squared value indicates the proportion of overall variation (55 %) in the data accounted for by the model. For a biological response, an R-square this small is not unusual.

Class	Levels	Values
Days	2	0 28
Conc	7	0 1 2.2 4.6 10 22 46

Label	Estimate	StdErr	DF	tValue	Probt
Growth: 46 mg/l vs 0	-0.9050	0.2803	117	-3.23	0.0016
Growth: 22 mg/l vs 0	-0.1656	0.2423	117	-0.68	0.4958
Growth: 10 mg/l vs 0	-0.1911	0.2423	117	-0.79	0.4319
Growth: 4.6 mg/l vs 0	-0.0800	0.2390	117	-0.33	0.7384
Growth: 2.2 mg/l vs 0	0.1022	0.2423	117	0.42	0.6739
Growth: 1 mg/l vs 0	0.0200	0.2390	117	0.08	0.9335

The appropriateness of this ANOVA analysis is assessed partly through the following test for normality. This test is done on the residuals from the ANOVA. As pointed out above, this test ignores the correlations of measurements on the same fish (since fish are not identified individually). The only significant comparison is at the 46 mg/l concentration. Accordingly, the NOEC is 22 mg/l, the same as by the Jonckheere-Terpstra test.

D.2.2.2.3 Shapiro-Wilk test of normality of length

SHAPIRO-WILK TEST OF NORMALITY OF length
Variable: Resid

Moments

N	131	Sum Weights	131
Mean	0	Sum Observations	0
Std Deviation	0.3585244	Variance	0.12853974
Skewness	-0.0318169	Kurtosis	-0.3018538

```
Uncorrected SS      16.7101667    Corrected SS         16.7101667
Coeff Variation              .    Std Error Mean       0.03132442

                    Basic Statistical Measures

          Location                      Variability
      Mean      0.00000     Std Deviation            0.35852
      Median   -0.01000     Variance                 0.12854
      Mode      0.38000     Range                    1.91500
                            Interquartile Range      0.51444

Fish Growth Example: Length
      Trout Size Data
       FULL DATA SET
                         Tests for Normality
      Test                   --Statistic---    -----p Value------
      Shapiro-Wilk           W      0.98607    Pr < W      0.2041
      Kolmogorov-Smirnov     D     0.065603    Pr > D      >0.1500
      Cramer-von Mises       W-Sq  0.080864    Pr > W-Sq   0.2078
      Anderson-Darling       A-Sq   0.59964    Pr > A-Sq   0.1199
```

This is default SAS Proc Univariate output, so four tests for normality are shown. It is highly recommended that each laboratoy select one of these tests and use it as *the* formal test of normality. Alternatively, a laboratory can develop a set of rules that indicate when to use which test. In either case, care should be taken so as not to be guilty of selecting the statistical result desired. In the present case, all tests yield the same conclusion. According to the Shapiro-Wilk test, the data are consistent with normality.

```
                    Variable:  Resid
                 Quantile      Estimate
                 100% Max      0.875000
                 99%           0.850000
                 95%           0.550000
                 90%           0.455556
                 75% Q3        0.270000
                 50% Median   -0.010000
                 25% Q1       -0.244444
                 10%          -0.510000
                 5%           -0.530000
                 1%           -0.640000
                 0% Min       -1.040000

          Stem Leaf                #             Boxplot
             8 58                  2                |
             7 8                   1                |
             6 166                 3                |
             5 15                  2                |
             4 2666677             7                |
             3 22666677788889     14                |
             2 1666677889         10             +-----+
             1 125568889           9             |     |
             0 12256666677788899  17             |  +  |
            -0 98844443322222211  17             *-----*
            -1 5444322221         10             |     |
            -2 4443322             7             +-----+
            -3 955544433           9                |
            -4 998443221           9                |
            -5 85433222221        11                |
            -6 44                  2                |
            -7
```

```
 -8
 -9
-10 4                        1                0
    ----+----+----+----+
Multiply Stem.Leaf by 10**-1
```

```
          Fish Growth Example: Length
                Trout Size Data
   SHAPIRO-WILK TEST OF NORMALITY OF length
                FULL DATA SET
               Variable:  Resid

                Normal Probability Plot
 0.85+                                                  *+++*
     |                                                +*+
     |                                             ***
     |                                          ++*
     |                                       ** *
     |                                    *****
     |                                 ****
     |                              +**
     |                           ****
     |                        ****
     |                     ***
     |                  +***
     |               ++****
     |             +*****
     |       *******
     |    ** +++
     |    +++
     |+++
     |
-1.05+*
      +----+----+----+----+----+----+----+----+----+----+
          -2        -1         0        +1        +2
```

D.2.2.2.4 Possible outliers from ANOVA on length

POSSIBLE OUTLIERS FROM ANOVA ON length

Obs	Days	Conc	Length	Pred	Resid	LB	UB
1	28	0	5.4	6.44000	-1.04	-1.01611	1.04167

D.2.2.2.5 Levene test for length

LEVENE TEST FOR length

Effect	DF	LEVENE	P_VALUE
Days*Conc	6	0.17219	0.98381

By Levene's test, there is no reason to reject the hypothesis that the within-group variances are equal. A standard ANOVA can be done.

D.2.3 Mass

D.2.3.1 Method 1

The Step-Down Jonckheere-Terpstra test is applied, first with all concentrations present, then with the 46 mg/l group omitted. The variable DeltaL is the difference of the day 28 mean minus the day 0 mean for each concentration. This was done using SAS Proc Freq. The default output includes both one- and two-sided tests for trend and both exact and asymptotic tests. These are all left for the reader to see, but only the one-sided

exact results are used in the discussion. It is observed, however, that for these data, the asymptotic and exact methods do not lead to the same NOEC.

Jonckheere-Terpstra Test of DeltaW Through Conc=46

Statistics for Table of DeltaW by Conc

Jonckheere-Terpstra Test	
Statistic (JT)	3.0000
Z	-2.2528
Asymptotic Test	
One-sided Pr < Z	0.0121
Two-sided Pr > |Z|	0.0243
Exact Test	
One-sided Pr <= JT	0,0151
Two-sided Pr >= |JT - Mean|	0,0302

Since the Jonckheere-Terpstra test with all concentrations included is significant at the 0,05 level, the high concentration is omitted and the test is repeated with the remaining concentrations.

Jonckheere-Terpstra Test of DeltaW Through Conc=22

Statistics for Table of DeltaW by Conc

Jonckheere-Terpstra Test	
Statistic (JT)	3.0000
Z	-1.6908
Asymptotic Test	
One-sided Pr < Z	0.0454
Two-sided Pr > |Z|	0.0909
Exact Test	
One-sided Pr <= JT	0.0681
Two-sided Pr >= |JT - Mean|	0.1361

The exact test is not significant at the 0,05 level, so no further testing is required and the NOEC is 22 mg/l. Note, however, that if standard asymptotic methods were used, the test would continue one step further, as shown below. There is no general expectation that asymptotic methods are more (or less) sensitive or powerful than exact methods. A laboratory should decide on rules for the use of exact and asymptotic methods and then use whatever conclusion follows.

Jonckheere-Terpstra Test of DeltaW Through Conc=10

Statistics for Table of DeltaW by Conc

Jonckheere-Terpstra Test	
Statistic (JT)	3.0000
Z	-0.9798
Asymptotic Test	
One-sided Pr < Z	0.1636
Two-sided Pr > |Z|	0.3272
Exact Test	
One-sided Pr <= JT	0.2417
Two-sided Pr >= |JT - Mean|	0.4833

D.2.3.2 Method 2

D.2.3.2.1 Full trout-size data set and basic statistics for mass

Fish Growth Example: Weight [24)]
Trout Size Data
FULL DATA SET

A two-factor ANOVA on length, with day and concentration, and their interaction as model terms. First, the simple means are computed. It is apparent that there is a dose response, perhaps beginning with the 2,2 mg/l concentration.

Obs	Conc	DeltaL	DeltaW
1	0	0.94000	1.20000
2	1	0.94000	1.10000
3	2.2	1.02222	1.27778
4	4.6	0.85000	1.17000
5	10	0.68889	0.97778
6	22	0.74444	0.93333
7	46	-0.07500	0.02500

Basic Statistics for weight

Days	Conc	Mean_ Weight	Std_weight	n_Weight
0	0	1.9	0.37712362	10
0	1	1.81	0.26436507	10
0	2.2	1.86	0.30983867	10
0	4.6	1.86	0.25473298	10
0	10	1.82	0.31198291	10
0	22	1.84	0.25905812	10
0	46	1.8	0.2981424	10
28	0	2.8	0.56764621	10
28	1	2.7	0.42687495	10
28	2.2	2.88	0.52387445	9
28	4.6	2.77	0.61472668	10
28	10	2.58	0.52387445	9
28	22	2.53	0.63245553	9
28	46	1.63	0.83416625	4

It is observed that mortality in the high concentration has noticeably affected the sample size on day 28. The other differences in sample size are quite small and are likely to have insignificant impact on conclusions.

CovParm	Estimate
Residual	0.1988

Effect	Num DF	Den DF	FValue	ProbF	MSERR	SSQRS	SSERR
Days	1	117	79.24	<.0001	0.19877	15.7509	23.2566
Conc	6	117	3.29	0.0050	0.19877	3.9250	23.2566
Days*Conc	6	117	2.62	0.0203	0.19877	3.1254	23.2566

24) The quantity “weight” is a force (gravitational force) and is measured in newtons (N). The quantity “mass” is measured in kilograms (kg). The use of the term “weight” in the extractions shown in the Courier New font listings is deprecated, but is retained in this Technical Specification as this is the way it appears in the actual extractions from the program(s) used.

These are the overall F-Tests for ANOVA. It is observed that both main effects and the interaction are significant at the 0,05 level. The CovParm above the F-tests is the pooled sample mean-squared error.

D.2.3.2.2 ANOVA summary statistics

ANOVA SUMMARY STATISTCS

MODELSS	SSERR	TOTSS	RSQUARE
26.1387	23.2566	49.3953	0.52917

These are basic measures that can be used for model assessment. Since this is a linear model, the R-squared value indicates the proportion of overall variation in the data accounted for by the model. For a biological response, an R-square this small is not unusual.

Fish Growth Example: Weight
Trout Size Data
CLASS LEVEL INFORMATION
FULL DATA SET

Class	Levels	Values
Days	2	0 28
Conc	7	0 1 2.2 4.6 10 22 46

Fish Growth Example: Weight
Trout Size Data
TESTS OF LINEAR CONTRASTS
FULL DATA SET

Label	Estimate	StdErr	DF	tValue	Probt
Growth: 46 mg/l vs 0	-1.0750	0.3306	117	-3.25	0.0015
Growth: 22 mg/l vs 0	-0.2067	0.2859	117	-0.72	0.4712
Growth: 10 mg/l vs 0	-0.1422	0.2859	117	-0.50	0.6198
Growth: 4.6 mg/l vs 0	0.01000	0.2820	117	0.04	0.9718
Growth: 2.2 mg/l vs 0	0.1178	0.2859	117	0.41	0.6811
Growth: 1 mg/l vs 0	-0.01000	0.2820	117	-0.04	0.9718

The appropriateness of this ANOVA analysis is assessed partly through the following test for normality. This test is done on the residuals from the ANOVA. As pointed out above, this test ignores the correlations of measurements on the same fish (since fish are not identified individually). The only significant comparison is at the 46 mg/l concentration. Accordingly, the NOEC is 22 mg/l, the same as by the exact Jonckheere-Terpstra test.

D.2.3.2.3 Shapiro-Wilk test of normality of mass

Fish Growth Example: Weight
Trout Size Data
SHAPIRO-WILK TEST OF NORMALITY OF Weight
FULL DATA SET

Variable: Resid
Moments

N	131	Sum Weights	131
Mean	0	Sum Observations	0
Std Deviation	0.42296218	Variance	0.17889701
Skewness	0.50903274	Kurtosis	0.53166703
Uncorrected SS	23.2566111	Corrected SS	23.2566111
Coeff Variation	.	Std Error Mean	0.03695438

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	0.42296
Median	-0.03333	Variance	0.17890
Mode	-0.73333	Range	2.23000
		Interquartile Range	0.54000

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.979035	Pr < W	0.0403
Kolmogorov-Smirnov	D	0.072519	Pr > D	0.0897
Cramer-von Mises	W-Sq	0.090825	Pr > W-Sq	0.1493
Anderson-Darling	A-Sq	0.632637	Pr > A-Sq	0.0978

The Shapiro-Wilk test is significant at the 0,05 level. A examination of the QQ-plot and stem-and-leaf plot below, as well as the p-value being just below 0,05, may suggest the violation of normality is minor. Also, while it is not included here to keep the text to a minimum, if the outliers identified below are omitted and the resulting dataset is re-analysed, the NOEC is the same and the significant Shapiro-Wilk test is eliminated. Altogether, this information suggests the present analysis can be accepted.

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.2300000
99%	1.1750000
95%	0.8000000
90%	0.5000000
75% Q3	0.2400000
50% Median	-0.0333333
25% Q1	-0.3000000
10%	-0.4777778
5%	-0.7000000
1%	-0.7777778
0% Min	-1.0000000

Fish Growth Example: Weight
Trout Size Data
FULL DATA SET

Variable: Resid

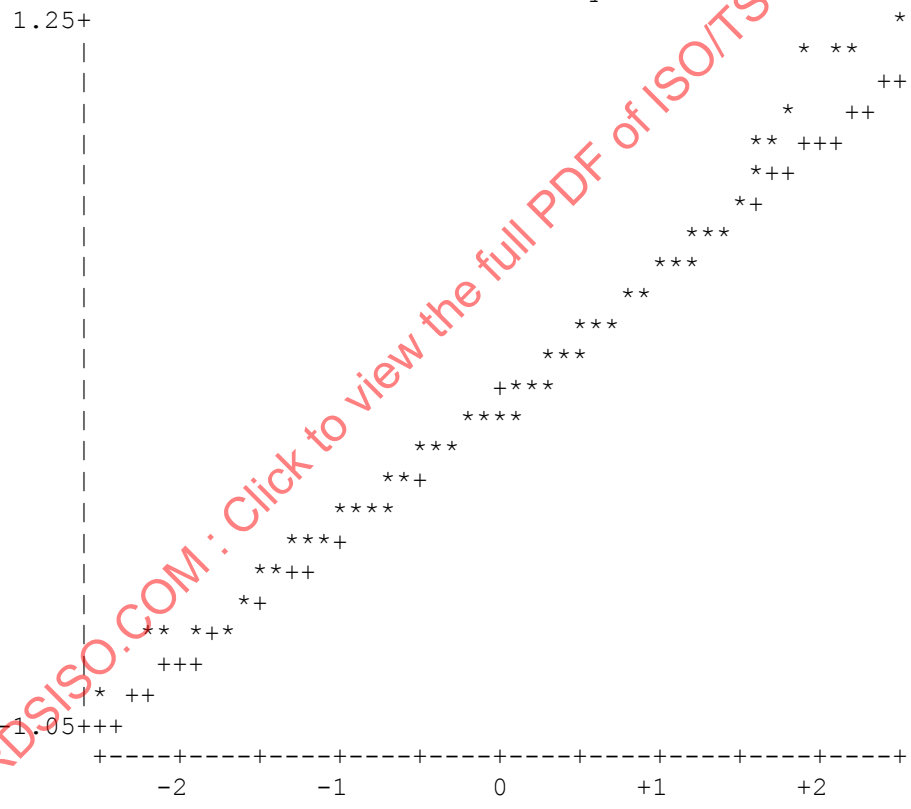
Stem	Leaf	#	Boxplot
12	3	1	0
11	028	3	0
10			
9	7	1	|
8	02	2	|
7	3	1	|
6	07	2	|
5	00247	5	|
4	000224	6	|
3	004668899	9	|
2	000234468	9	+-----+
1	0002344789	10	| |
0	000002444689	12	| + |
-0	776664432211	12	*-----*
-1	88866644200000	14	| |
-2	8881100000	10	| |
-3	8766644321100	13	+-----+
-4	87622200	8	|

```
 -5 7300                    4                |
 -6 80                      2                |
 -7 873320                  6                |
 -8                                          |
 -9                                          |
-10 0                       1                |
    ----+----+----+----+
Multiply Stem.Leaf by 10**-1
```

```
          Fish Growth Example: Weight
                Trout Size Data
   SHAPIRO-WILK TEST OF NORMALITY OF Weight
                FULL DATA SET

             The UNIVARIATE Procedure
                 Variable:  Resid

              Normal Probability Plot
```



D.2.3.2.4 Possible outliers from ANOVA on mass

```
          Fish Growth Example: Weight
                Trout Size Data
   POSSIBLE OUTLIERS FROM ANOVA ON Weight
                FULL DATA SET
```

Obs	Days	Conc	Weight	Pred	Resid	LB	UB
1	28	0	3.9	2.80000	1.10000	-1.11	1.05
2	28	2.2	4	2.87778	1.12222	-1.11	1.05
3	28	4.6	4	2.77000	1.23000	-1.11	1.05
4	28	46	2.8	1.62500	1.17500	-1.11	1.05

D.2.3.2.5 Levene test for mass

LEVENE TEST FOR Weight

Effect	DF	LEVENE	P_VALUE
Days*Conc	6	0.46436	0.83347

By Levene's test, there is no reason to reject the hypothesis that the within-group variances are equal. A standard ANOVA can be done.

D.3 Example of data analysis by dose-response modelling

D.3.1 General

It is assumed that an EC_{10} is required.

According to the flow-chart of Clause 6, the dose-response data should include more than two concentration groups with different response levels. The dose-response data of this data set (for both lengths and masses) do not clearly comply with this requirement (see figures below). However, in view of the relatively large number of concentration groups, and the general trend of decreasing response with concentration, the data might nonetheless be suitable for deriving an EC_{10}. In accordance with the general recommendation of Clause 6, various models are fitted to the data.

D.3.2 Dose-response analysis for fish mass

In fitting the dose-response model, the mass data are assumed to be log-normally distributed. Hence, the model is fitted to the data on log-scale, i.e. both the data and the model prediction is log-transformed. Note, however, that the models used describe the response variable on the original scale as a function of dose.

First the nested non-linear model proposed by Slob (2002) is fitted. This results in the following log-likelihoods:

Model 1	$y = a$	loglik = –5,21
Model 2	$y = a \exp(x/b)$	loglik = 4,85
Model 3	$y = a \exp[\pm(x/b)^d]$	loglik = 6,53
Model 4	$y = a\,[c - (c - 1) \exp(-x/b)]$	loglik = 4,85
Model 5	$y = a\,\{c - (c - 1) \exp[-(x/b)^d]\}$	loglik = 6,53

The log-likelihoods can be compared with the likelihood-ratio test. A model with one more parameter fits the data significantly better (at $\alpha = 0{,}05$) than the model without that parameter when the increase in the log-likelihood is greater than 1,92. The difference in log-likelihoods between Model 2 and Model 1 is 10,06, so there is no doubt that the data show a significant dose-response. While Model 3 results in a higher log-likelihood than Model 2, the difference is not significant, and it may be concluded that Model 2 is from this family of models the appropriate one for describing the data. Figure D.1 shows the results for this model.

Next a polynomial model is fitted to these data. Since this is again a nested family of models, the log-likelihoods can be compared by the ratio-likelihood test:

Model 1	$y = a$	loglik = –5,21
Model 2	$y = a + bx$	loglik = 5,55
Model 3	$y = a + bx + cx^2$	loglik = 6,58
Model 4	$y = a + bx + cx^2 + dx^3$	loglik = 6,69

Here, Model 2 (straight line) is not significantly improved by higher-order polynomials, and this model may be selected from this family of models.

Finally, the power model, $y = c + ax^b$, is fitted to the masses. This model results in a log-likelihood value of 6,57. However, by fixing the parameter b to unity, the model reduces to a straight line. Hence, the power model and the straight line are nested, and their log-likelihoods can be compared. Since the straight line resulted in a log-likelihood of 5,55, the power model does not give a significantly better fit.

None of the three-parameter models gives a significantly better fit than its two-parameter counterpart. As Table D.4 shows, both two-parameter models give similar results. The table also shows the results for two models with three parameters, even though they did not result in a significantly better fit than the associated two-parameter model. As discussed in Clause 6, a model that contains too many parameters (i.e. in relation to the data) may results in wider confidence intervals for the EC_{10}. This is clearly illustrated in this case.

It may be concluded that both two-parameter models give similar results. Also, the confidence intervals are similar. Hence, although this particular data set only shows two clearly different response levels, the data apparently contain sufficient information to warrant the estimation of the EC_{10}. This may be explained by the relatively large number of concentrations tested, most of which show no or only a small response. Thus, the estimated EC_{10} is sufficiently supported by surrounding concentrations that have been tested.

Table D.4 — Summary of results for exponential and straight-line models

Model [a]	Log-lik	EC_{10}	90 % Confidence interval
$y = a \exp(x/b)$ [b]	4,85	9,19	6,86 to 13,94
$y = a \exp(x/b^c)$ [c]	6,53	21,08	9,89 to 41,39
$y = \quad + bx$ [d]	5,55	10,55	8,56 to 14,25
$y = a + bx^c$ [e]	6,57	20,70	8,65 to 34,34

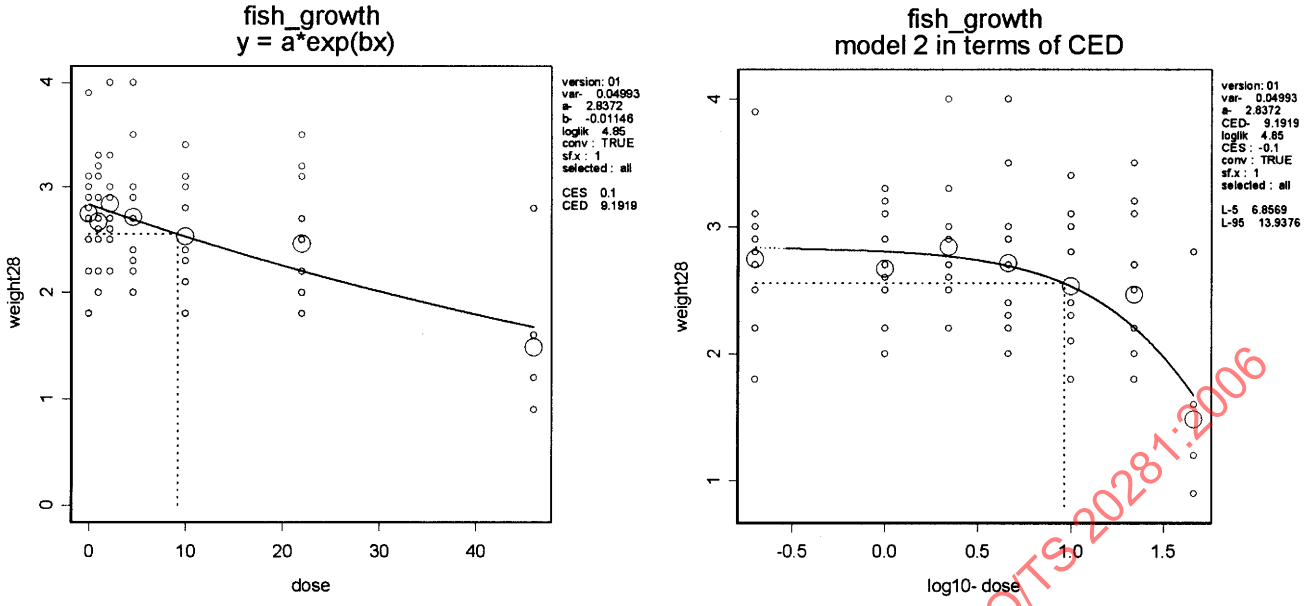
a Note that both the first two and the second two models form a nested couple.

b Exponential model, with two parameters.

c Exponential model, with three parameters.

d Straight-line model, with two parameters.

e Straight-line model, with three parameters.



a) Concentration on x-axis

b) log10-concentration on x-axis (to improve visibility)

$CED = EC_{10}$.

NOTE The model was re-parameterized here to make the EC_{10} a parameter in the model (instead of b). In this way, the confidence interval (L – 5 to L – 95) for the EC_{10} can be calculated by the likelihood profile method.

Large marks: geometric means

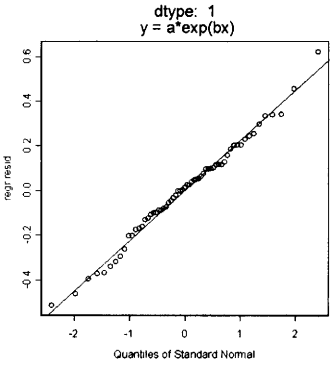
Figure D.1 — Exponential model, $y = a \exp(bx)$, fitted to masses at 28 days

Assumptions

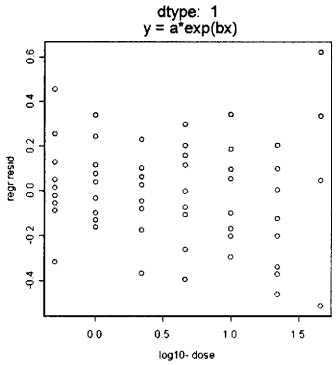
To check the assumptions of normality and homogeneous variances, the regression residuals (i.e. the deviations of the individual data points from the dose-response model) may be plotted in various ways. Here, two plots are considered. One is the so-called QQ-plot, where the observed quantiles are plotted against the theoretical quantiles, e.g. according to the normal distribution. When data are sampled from a normal distribution, this plot should theoretically result in a straight line. It should be noticed that fitting a line to a QQ-plot is unsound (which is not always recognized). One may draw the theoretical straight line in the plot, with intercept equal to the mean of the data points and with slope equal to the standard deviation of the data points. In the case of regression, the data points are the regression residuals, which are corrected for the dose-response relationship.

In interpreting a QQ-plot one should realize that, due to sampling errors, fluctuations around the line can easily arise, especially in small data sets. In particular, a pattern resembling Aesculapius' staff is not unusual, even for data that are sampled from a normal distribution by the computer. Hence, QQ-plots should only lead to the conclusion that the assumed distribution is inadequate when the data show a clear overall curvature. It is always the general trend, not single data points that should be considered.

In Figure D.2, the regression residuals resulting from the analysis on log-scale are plotted. As the left panel shows, the data did comply with the assumption of log-normality. The residuals plotted against concentration do not reveal a clear trend, and the assumption of homogeneous variances (on log-scale) appears acceptable (see right panel of Figure D.2).



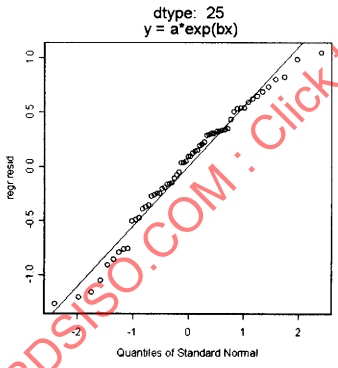
a) QQ-plot (for normal distribution) of the regression residuals for the analysis of Figure D.1, where the model was fitted on log-scale



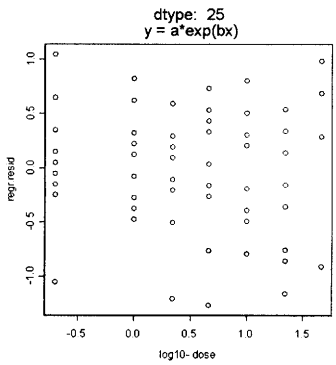
b) The same residuals plotted against dose (dose on log-scale to improve visibility

Figure D.2 — Plots of regression residuals for the analysis of Figure D.1

Although not strictly needed, the residual plots are also shown for an analysis where the log-transformation was omitted (see Figure D.3). The QQ-plot now shows a slight curvature, which confirms the use of a log-transformation. It should be noted that the scatter in these data is relatively mild (CV = 23 %), and that a log-normal distribution gets closer to a normal distribution with smaller variation (CV). Therefore, the smaller the scatter in the data, the more data are needed to see any difference in the QQ-plots assuming normality or log-normality. For the same reason, it may be expected that applying or omitting the log-transformation has no large impact on the results of the analysis when the scatter in the data is relatively small.



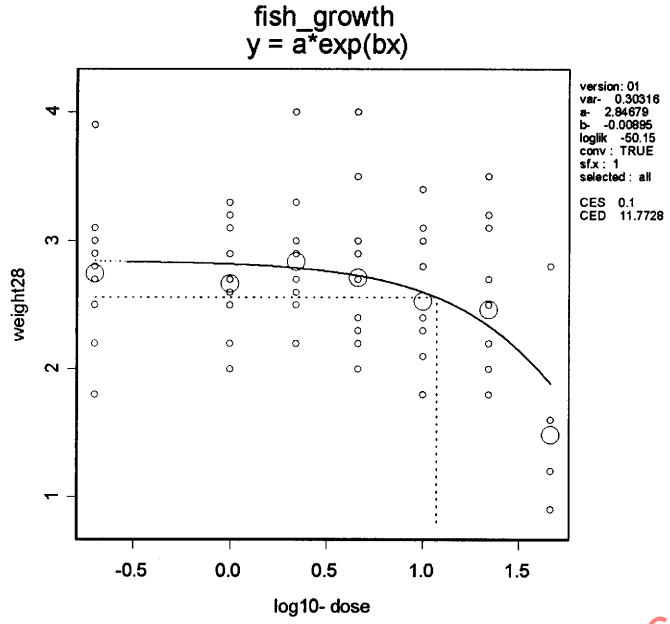
a) QQ-plot (normal distribution) of the regression residuals for an analysis as in Figure D.1, where the model was fitted without transformation



b) The same residuals plotted against dose (dose on log-scale to improve visibility)

Figure D.3 — Plots of regression residuals

The analysis shown in Figure D.1 was repeated but now omitting the log-transformation (see Figure D.4). The results are somewhat different, especially with regard to the upper confidence limit for the EC_{10} (see Table D.5).



NOTE Large marks: arithmetic means.

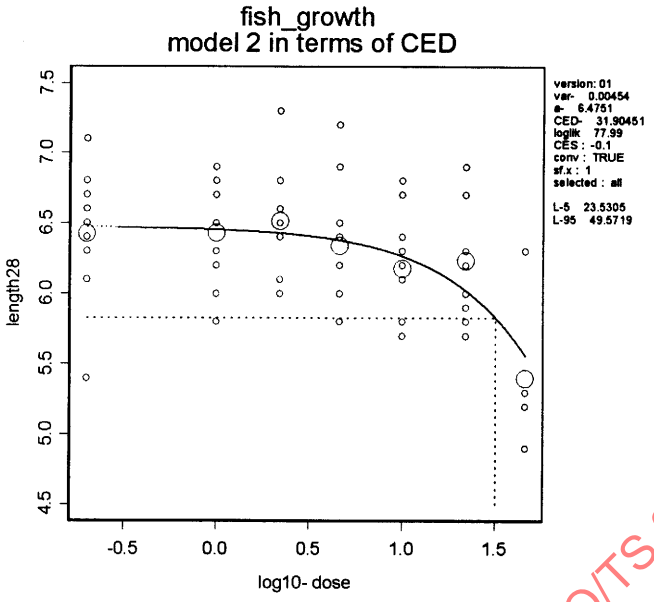
Figure D.4 — Dose-response analysis of the fish masses, but without log-transformation

Table D.5 — Summary of results for the exponential model fitted to the log-transformed masses, and the untransformed masses

	EC_{10}	90 % Confidence interval
Analysis with log-transformation	9,19	6,86 to 13,94 [a]
Analysis without transformation	11,78	7,79 to 21,94 [a]
[a] Obtained by the likelihood profile method.		

D.3.3 Dose-response analysis for fish length

The analysis of the length data is very similar to that of the mass data. Figure D.5 shows the exponential model fitted to the lengths (assuming lognormality). The EC_{10} is estimated at 31,90 mg/l, as compared to 9,19 mg/l for the masses (same model fitted). Therefore, the mass data would most likely be used for deriving an EC_{10}. However, one might argue that a decrease in body length by 10 % is not equivalent to a decrease in body mass by 10 % from a biological point of view.



NOTE Large marks: geometric means.

Figure D.5 — Exponential model fitted to the body lengths

D.4 Examples of data analysis using DEBtox (biological methods)

D.4.1 Data

The mean initial volumetric length of *Oncorhynchus mykiss* is 1,222 $g^{1/3}$.

Table D.6 — Response versus concentration at 21 d

Concentration in mg/l	0	1	2,2	4,6	10	22	46
Response in mean volumetric length, $g^{1/3}$	1,403	1,389	1,418	1,398	1,365	1,355	1,152

D.4.2 Parameter estimates and asymptotic standard deviations (ASD)

No effect concentration	5.597	mg l^{-1}	7.399		
Blank ultimate length	15.84	$g^{1/3}$	1.221	-0.456	
Tolerance concentration	43.78	mg l^{-1}	11.884	-0.818	0.284
Elimination rate	Infinity	d^{-1}			
Initial length	1.222	$g^{1/3}$			
Von Bertalanffy growth rate	0.00059	d^{-1}			
Energy investment ratio	1				
Mean deviation	0.03006	$g^{1/3}$			

Figure D.6 — DEBtox example: Parameter estimates and asymptotic standard deviations (ASD)

Table D.7 — EC_x values (derived from parameter values) in milligrams per litre

Day	EC_0	ASD	EC_{10}	ASD
21	5,6	7,4	37,4	5,71

D.4.3 Graphical test of model predictions against data

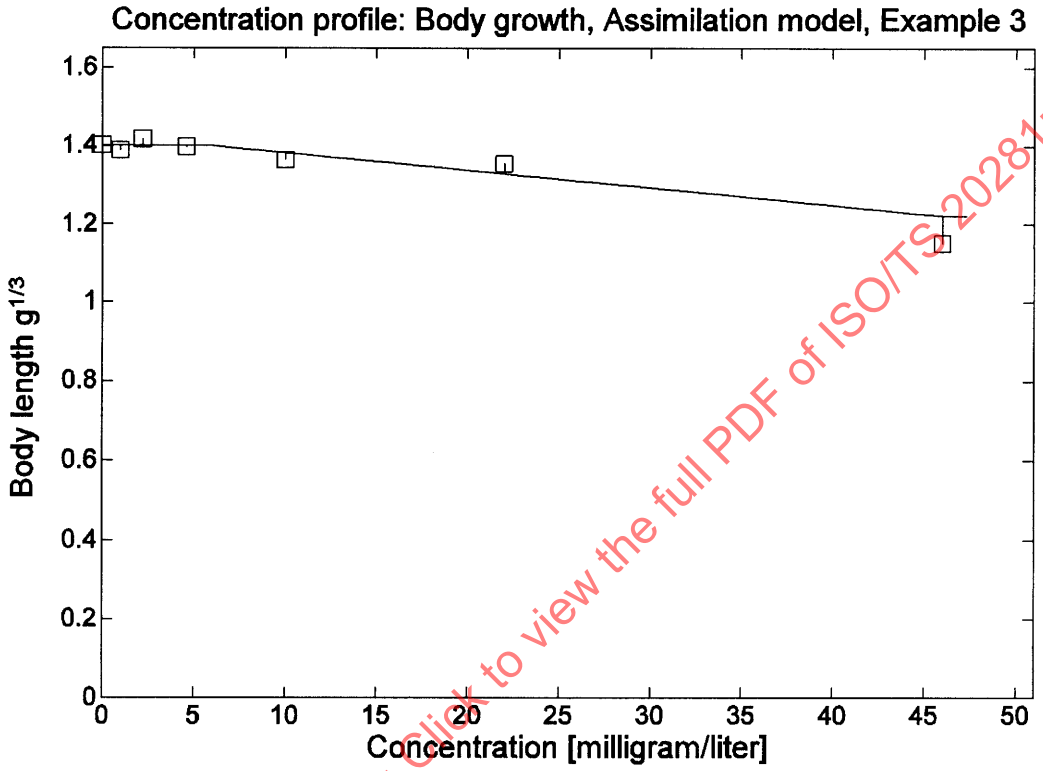


Figure D.7 — DEBtox example: Concentration profile

D.4.4 Profile likelihood for NEC estimate

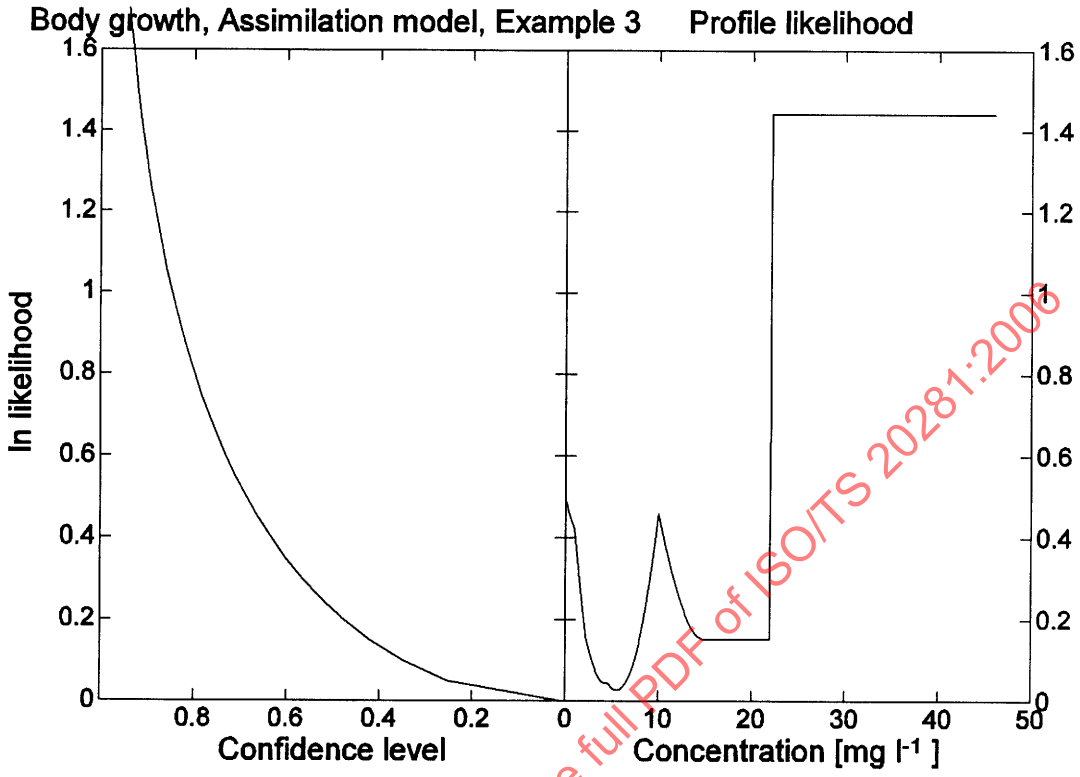


Figure D.8 — DEBtox example: Profile likelihood for NEC estimate

D.4.5 Comments

The mean body size at the highest concentration was lower than the initial one; the increase in mass during 21 d in the blank was very small. The model for effects on assimilation fitted slightly better than for effects on maintenance or growth costs. The von Bertalanffy growth rate was fixed at 5,9 10^{-4} d (from Kooijman, 2000), and the initial size at 1,222 $g^{1/3}$ (measured mean value). The EC_{50} is on body length, which is not meaningful in this case since 50 % of the blank length is far below the initial length. The NEC value does not differ significantly from zero, with a 95 % confidence interval of 0 mg/l to 22 mg/l.

Annex E
(informative)

Description and power of selected tests and methods

E.1 Description of selected methods for use with quantal data

E.1.1 Cochran-Armitage trend test

E.1.1.1 General

Quantal (binary) data can be collected and categorized by explanatory factors (such as dosage or treatment level). An analysis of such data usually tries to indicate relationships between the response (binary) variable and factors such as dose level. In such cases, the Pearson Chi-squared (χ^2) test for independence can be used to find if any relationships exist. The Cochran-Armitage test decomposes the Pearson Chi-squared test into a test for linear trend for the dose-response and a measure of lack of monotonicity,

$$\chi^2_{(k-1)} = \chi^2_{(1)} + \chi^2_{(k-2)}$$

where

$\chi^2_{(1)}$ is the 1 d*f* calculated Cochran-Armitage linear trend statistic; and

$\chi^2_{(k-2)}$ is the k–2 d*f* Chi-squared test statistic for lack of monotonicity.

Suppose the number of affected individuals is Y_i within a group of N_i animals exposed to dose X_i.

The proportion affected by dosage X_i is $p_i = Y_i/N_i$.

The model is

$$p_i = H(\alpha + \beta z_i),$$

where

z_i is the dose-metric (e.g. log-dose or dose rank); and

H is some twice-differentiable, monotone link function such as the logistic.

The test for linear trend is a test for $\beta = 0$.

Standard weighted regression gives the estimate of β as

$$b = \frac{\sum_{i=1}^{k} n_i (p_i - \bar{p})(z_i - \bar{z})}{\sum_{i=1}^{k} n_i (z_i - \bar{z})^2},$$

where

$\bar{z}$ is the weighted mean of the z_i;

$\bar{p}$ is the weighted sum of the p_i.

The test statistic for $\beta = 0$ is b^2/ Var(b) which is approximately distributed as $\chi^2_{(1)}$. Formally written, the Cochran-Armitage Chi-squared is

$$\chi_1^2 = \frac{\left(\sum_{i=1}^{k} y_i z_i - \frac{T_y \left(\sum_{i=1}^{k} N_i z_i \right)}{T} \right)^2}{\overline{pq} \left(\sum_{i=1}^{k} N_i z_i^2 - \frac{\left(\sum_{i=1}^{k} N_i z_i \right)^2}{T} \right)}$$

where

$T_y = \Sigma\ Y_i$;

$T = \Sigma\ N_i$;

$\bar{q}$ is the weighted sum of the $q_i = 1 - p_i$.

The Cochran-Armitage Chi-squared can also be expressed as a z-statistic in order to take account of the direction of the trend. The z-statistic is obtained from the formula given by removing the exponent 2 in the numerator and taking the square-root of the denominator. This z-test has a standard normal distribution under the null hypothesis of no trend, and the probability of the z-statistic can be obtained from a table of areas under the standard normal distribution. Only the z-statistic is appropriate for one-sided tests. Unlike the one-sided test, the χ^2 version of the Cochran-Armitage test can remain significant in a step-down application even when there is a change in direction of the trend. To avoid this situation when doing a two-sided test, one applies both one-sided z-tests with all doses present at the $\alpha/2$ level. At most one of these can be significant. If one is significant, this determines the direction of the trend and all further tests are done with the z-statistic for that same direction at the $\alpha/2$ level.

A general linear trend model for quantal data is

$p_i = H(a + bd_i)$,

where

a and b are parameters to be estimated;

d_i is some metric (measure) of the exposure level (e.g. dose, concentration, log concentration);

p_i is the probability of response; and

H is some monotone function, (referred to as a link function), e.g.

logistic,	$H_1(u) = e^u/(1 + e^u)$,
probit,	$H_2(u) = M(u)$,
extreme value,	$H_3(u) = 1 - \exp(e^u)$,
one-hit,	$H_4(u) = 1 - e^{-u}$.

For example, the trend model for the one-hit link function could be written as

$$p_i = 1 - \exp[-(a + bd_i)].$$

Tarone and Gart (1980) showed that for any link function likely to be of practical use, the same test statistic for significance of trend always arises from likelihood-type considerations, namely, the Cochran-Armitage test. Thus, it is not necessary to postulate the particular form of the link function. They also showed that this test is in general the most efficient test of trend for any monotone model. This is one of the reasons some regard the Cochran-Armitage test as inherently non-parametric. Hirji and Tang (1998) discuss the favourable power properties of the Cochran-Armitage test compared to various alternatives. They offer evidence that this test can be used for small samples and sparse data without undue concern that it is an asymptotic test.

E.1.1.2 Assumptions

Subjects are independent within and among groups and subgroups (if present). Group proportions are monotonic with respect to the dose score. While formally, this is a test for linear trend in the response in relation to the dose score used, it is generally the most powerful test against any monotone dose-response alternative. Furthermore, rank-order of doses (or equally-spaced dose scores) is used to reduce dose-dependence in the test. Note that robust versions of this test are available which allow for extra-binomial variation, notably one based on a beta-binomial model.

E.1.1.3 Power

Extensive power simulations of the step-down Cochran-Armitage test have demonstrated that in every instance considered where there is a monotone dose-response, the step-down application of the Cochran-Armitage test is more powerful than Fisher's Exact test. These simulations followed the step-down process to the NOEC determination, and covered a range of dose-response shapes, thresholds, background rates, number of treatment groups and number of subjects per group. These simulation results are useful in design of experiments and are being prepared for publication by J.W. Green. A small sample giving an idea of the results is found in E.2. These should be useful in experimental design.

E.1.1.4 Confidence intervals for incidence rate

Given the discussion above of Tarone and Gart (1980), confidence intervals for the true incidence rate at a given tested concentration can be calculated from one of the regression models described in Clause 6 for quantal data. There is no direct link between the Cochran-Armitage test statistic and these confidence intervals.

E.1.1.5 Example from a trout early life stage experiment

E.1.1.5.1 Data set

The data are from a trout early life stage experiment. Initially, there were 20 eggs placed in each of four replicate subgroups per concentration. Of those eggs that hatched (shown as the *Number at risk* value), some larvae did not survive to the time of thinning. The analysis is of the larval survival to time of thinning. There were two control groups, one water-only and the other including a solvent that is also present in all positive dose groups. There were no mortalities in either control, so they were combined for further analysis.

Dose (PPM)	Number at risk	Number Responding	% Responding	Dose Score
0	125	0	0.0	1
1	62	1	1.6	2
2	62	1	1.6	3
4	60	2	3.3	4
8	65	0	0.0	5
16	72	10	13.9	6
32	65	29	44.6	7

E.1.1.5.2 Cochran-Armitage test using equally-spaced dose scores

Cochran-Armitage test is one-sided for INCREASE in RESPONSE.

All doses included:

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	140.23874	6	0	
Trend	8.7567722	1	0	**
LOF	63.55768	5	2.231E-12	

With the 32 ppm concentration omitted:

SOURCE	Test_stat	Deg_Free	*p*-value	SIGNIF
Overall	34.479817	5	1.9107E-6	
Trend	4.1393854	1	0.0000174	**
LOF	17.345306	4	0.001656	

With the 32 ppm and 16 ppm concentrations omitted:

SOURCE	Test_stat	Deg_Free	p-value	SIGNIF
Overall	5.3062133	4	0.2572958	
Trend	0.7720901	1	0.2200305	
LOF	4.7100902	3	0.1942989	

No further testing is required. NOEL is current high concentration, 8 mg/l.

E.1.1.5.3 Subgroup proportions

E.1.1.5.3.1 Analysis of subgroup proportions — General

There were actually four replicate subgroups in each concentration in this experiment, which this analysis did not take into account. The data, broken down by replicate subgroup, are as follows. With the possible exception of the high concentration, there is no significant evidence of extra binomial variation in these data. While it is clear that by any reasonable analysis, the high dose group is an effects level, and there is no reason evident in the data to suggest a need to model replicate subgroups, it may be illustrative to do such an analysis.

For this purpose, we analyse subgroup proportions, which are reported below. With the large number of zeros, an analysis based on the normality of the proportion dead, even after an arc-sine square-root or Freeman-Tukey transformation is not justified. Rather, a Jonckheere-Terpstra test is done. While this ignores differences in sample sizes among the subgroups, these differences are small, so little precision is lost by such an analysis. Both the Jonckheere-Terpstra and Dunn tests find the NOEC to be dose 5, or 8 mg/l, the same as the Cochran-Armitage test.

E.1.1.5.3.2 Dunn's multiple comparisons (increasing) on DEADPROP

DOSE	COUNT	N0	MRANK	ABS_DIFF	CRIT05	CRIT01	SIGNIF	P_VAL
0	8	8	10.500	0.000	9.7600	11.9665		.
1	4	8	13.250	2.750	11.9535	14.6559		1.000
2	4	8	13.750	3.250	11.9535	14.6559		1.000
4	4	8	16.500	6.000	11.9535	14.6559		0.688
8	4	8	10.500	0.000	11.9535	14.6559		1.000
16	4	8	26.625	16.125	11.9535	14.6559	**	0.004
32	4	8	30.375	19.875	11.9535	14.6559	**	0.000

E.1.1.5.3.3 Step-down Jonckheere-Terpstra test

MONOTONICITY CHECK OF DEADPROP

PARM	P_T	SIGNIF
DOSE TREND	0.0001	**
DOSE QUAD	0.0001	**

KEY

ZC IS JONCKHEERE STATISTIC COMPUTED WITH TIE CORRECTION

ZCCF IS ZC WITH CONTINUITY CORRECTION FACTOR

P1UPCF IS *p*-VALUE FOR UPWARD TREND

P1DNCF IS *p*-VALUE FOR DOWNWARD TREND

p-VALUES ARE FOR TIE-CORRECTED TEST WITH CONTINUITY CORRECTION FACTOR

SIGNIF RESULTS ARE FOR AN INCREASING ALTERNATIVE HYPOTHESIS

Hi_Dose	JONC	ZC	ZCCF	P1UPCF	P1DNCF	SIGNIF
32	332	4.4091617	4.4281667	4.7519E-6	0.9999943	**
16	220.5	3.0749658	3.1003788	0.0009664	0.9988541	**
8	124.5	0.9908917	1.0305274	0.1513813	0.8292628	

E.1.1.5.3.4 Raw data for Dunn and Jonckheere-Terpstra examples

TOTRISK is the number of larvae exposed.

DEADLARV is the number of exposed larvae found dead by the end of the experimental period.

PROPDEAD = DEADLARV/TOTRISK.

REP is an identifier for a replicate subgroup within a given concentration (CONC).

CONC	REP	TOTRISK	DEADLARV	PROPDEAD
0	1	16	0	0
0	2	16	0	0
0	3	16	0	0
0	4	15	0	0
0	5	16	0	0
0	6	15	0	0
0	7	15	0	0
0	8	16	0	0
1	1	16	1	.0625
1	2	15	0	0
1	3	15	0	0
1	4	16	0	0
2	1	16	0	0
2	2	15	1	.065
2	3	16	0	0
2	4	15	0	0
4	1	14	0	0
4	2	15	1	.065
4	3	15	0	0
4	4	16	1	.065
8	1	15	0	0
8	2	16	0	0
8	3	17	0	0
8	4	17	0	0
16	1	17	2	.1176
16	2	17	3	.1765
16	3	18	2	.1111
16	4	20	3	.15
32	1	15	8	.5333
32	2	17	3	.1765
32	3	15	8	.5333
32	4	18	10	.5556

E.1.2 Fisher's Exact test

E.1.2.1 General

Fisher's Exact test is based on a 2×2 contingency table where control and a single treatment group are compared according to their prospective counts (affected/not affected). The diagram below illustrates this case.

Table E.1 — Fisher's Exact test: Contingency table

	Control group	Treatment group	Total
Affected	n_{00}	n_{01}	$n_{0.}$
Not affected	n_{10}	n_{11}	$n_{1.}$
Total	$n_{.0}$	$n_{.1}$	$n_{..}$

Fisher's Exact test is based on the probability of observing n_{01} affected subjects in the treatment group, if all marginal totals are considered fixed. This probability is given by the hypergeometric distribution: the stated probability of observing n_{01} affected subjects in the treatment group, given that $n_{0.}$ subjects were affected overall and a total of $n_{..}$ subjects were in both groups combined is

$$P(x = n_{01}) = \frac{\binom{n_{.0}}{n_{0.} - n_{01}}\binom{n_{.1}}{n_{01}}}{\binom{n_{..}}{n_{0.}}}.$$

From this, of course, the probability of observing at least n_{01} subjects in the treatment group is

$$P(x \geq n_{01}) = \sum_{i=n_{01}}^{n_{0.}} P(X = i).$$

This gives the significance level of the observed response for a one-sided test of an increase in the treatment group. Fisher's Exact test can be applied to compare each treatment group to the control, independently of all other treatment to control comparisons. When this is done, a Bonferroni-Holm adjustment for the number of comparisons being made can be applied to control the over-all false positive rate.

E.1.2.2 Power

The power of Fisher's Exact test is available in several software packages, including StatXact 4, PASS, and Study Size. The second of these can be found at http://www.ncss.com. The third is available at StudySize@CreoStat.com [25]. A simple search of the internet locates several such sites, some of which have free downloads. It is important to understand that the power of Fisher's Exact test depends on the background (i.e. control) incidence rate, with power decreasing as background rate increases (up to 50 %), when all other factors are fixed.

E.1.2.3 Assumptions

Subjects are independent within and among groups and subgroups (if present).

25) StatXact, PASS, and Study Size are examples of suitable products available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of these products.

E.1.2.4 Confidence intervals

Since no dose response relationship is assumed in using Fisher's Exact test, only a crude confidence interval based on the binomial distribution or its normal approximation can be constructed for the incidence rate at a tested concentration.

E.1.2.5 Example

The same data are analysed as for the Cochran-Armitage example. In this instance, Fisher's Exact test reports the same NOEC as the tests given above.

FISHER EXACT TEST vs CONTROL FOR DEADLARV OBSERVED

TESTING FOR AN INCREASING ALTERNATIVE HYPOTHESIS

Dose	Number at risk	Number Responding	*p*-value (Right)	Significance
1	62	1	0.33155	
2	62	1	0.33155	
4	60	2	0.10400	
8	65	0	1.00000	
16	72	10	0.00003	**
32	65	29	0.00000	**

E.1.3 Poisson tests

E.1.3.1 General

Similar to the context of the Cochran-Armitage test, quantal (binary) data can be collected and categorized by explanatory factors (such as dosage or treatment level). An analysis of such data usually tries to indicate relationships between the response (binary) variable and factors. The counts within subgroups are assumed to follow a Poisson distribution whose mean is a function, usually linear, of some dose metric. The use of equally spaced dose scores (i.e. rank-order) makes this more a test for trend than raw doses do, when there are large differences in doses. An advantage of the Poisson model over the Jonckheere for quantal data is the use of a rate multiplier to adjust for unequal sample sizes.

Suppose the number of affected individuals is Y_{ij} within subgroup j of group i, a group of n_{ij} animals exposed to the ith dose with dose score z_i. The proportion affected by this dosage is $p_i = Y_i/n_{ij}$. The model is determined by

$$p[Y_{ij} = y] = \frac{e^{-\mu_i}\mu_i^y}{y!},$$

where μ_i is a function of dose.

E.1.3.2 Trend version of the Poisson test

For the trend version of the Poisson test, it is typical to model

$$\log(\mu_{ij}) = \log(r_{ij}) + \alpha + \beta z_i,$$

where z_i is the dose-metric (e.g. log-dose or dose rank), r_{ij} is a rate multiplier (which can be the subgroup size) for the jth subgroup of the ith group, and α and β are parameters to be estimated. The test for linear trend is a test for $\beta = 0$.

For the Poisson trend model, one fits the model with all dose groups present and tests the hypothesis $\beta = 0$ against the alternative $\beta > 0$. If this hypothesis is rejected at the 0,05 significance level, then the high dose group is omitted and the model is re-fit to the remaining dose groups. This process is continued until the hypothesis of positive slope is first not rejected. The high dose remaining at this stage is the NOEC.

E.1.3.3 Non-trend version of the Poisson test

For a non-trend version of this test, βz_i is replaced by $z_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a parameter vector $<\beta_l>$ and z_i is a column vector $<z_{iu}>$ whose uth component is zero unless $u = i$. For $i = 0$ (i.e. for the control group), all components of z_i are zero.

One then tests the hypotheses $\beta_i = 0$, for $i = 1, 2, \ldots$ against the obvious one- or two-sided alternatives. Generally, a Bonferroni-Holm adjustment is made for the number of such comparisons made.

E.1.3.4 Other tests for lack of fit of the Poisson model

There are tests for lack of fit of the Poisson model, which are not described here.

References include McCullagh and Nelder (1989), Collett (1991), Aitkin *et al*. (1989), Morgan (1992), Mehta and Patel (1999), Hosmer and Lemeshow (1989), Thomas (1983).

Software includes the GENMOD procedure in SAS version 8 and LogXact 4[26)] for Windows. The availability in LogXact of exact permutation procedures is especially appropriate for small samples or rare events.

Robust versions of the Poisson tests are available. [Weller and Ryan (1998); Hirji and Tang (1998); Breslow (1990); Tarone and Gart (1980)].

E.1.3.5 Assumptions

Subjects are independent within and among groups and subgroups (if present). Within-group counts follow a Poisson distribution.

— For the trend version, a monotone trend in the dose-response is assumed.

— For the non-trend version, the dose metric is not used. Rather, an ANOVA-type model is used which assumes Poisson distribution rather than normal.

Note that robust versions of this test are available which allow for extra-binomial variation among subgroups.

26) GENMOD and LogXact are examples of suitable products available commercially. This information is given for the convenience of users of this Technical Specification and does not constitute an endorsement by ISO of these products.

E.2 Power of the Cochran-Armitage test

E.2.1 General

It is important to understand that the power of the Cochran-Armitage test depends on the background (i.e. control) incidence rate, with power decreasing as background rate increases (up to 50 %), when all other factors are fixed.

The powers associated with the step-down application of the Cochran-Armitage test are not available, so far as we know, in any commercial or on-line source or even in published form. For that reason, a set of power plots has been included. The power of a statistical test also depends on the size effect to be found, the number of observations per treatment group and other factors. In comparing percent effect of treatments to a common control using a step-down trend test, additional factors affecting power are the shape of the concentration-response curve, the number of concentrations and the true threshold of toxicity (if such a thing exists). This last point is relevant only in that it affects the concentration-response shape by affecting the concentration at which the concentration-response begins to depart from horizontal.

A full treatment of the power of the step-down Cochran-Armitage test is being prepared for publication. As an aid in designing experiments, the following power curves can be used. These are for a response following a linear concentration-response shape, with no lag (i.e. threshold = 0), for sample sizes 20, 40, 60 and 80 and number of concentrations ranging from 3 to 5.

A further consideration is whether the test is done in a one-sided or two-sided fashion. For quantal responses, it is unusual to be interested in anything but an increased incidence rate, so only one-sided powers are presented. Powers for a two-sided test are, of course, lower for the same set of conditions.

E.2.2 Interpretation of the power curves

Three plots are presented for each of the sample sizes 20, 40, 60 and 80 subjects per concentration. The first of a set of three plots show the power of detecting an effect of a given size in the highest dose. The second of a set of three plots shows the power of detecting an effect of a given size at stage two of the step-down process; that is, in the second highest dose. The third of a set of three plots shows the power of detecting an effect of a given size in the third highest dose, that is, at stage three of the step-down process. For each plot, the vertical axis is the power or probability (expressed as a percent) of finding a significant effect if the true effect has the magnitude given on the horizontal axis. The horizontal axis shows the true change in percent effect from the control or background rate. Five power curves are drawn in each plot corresponding to background rates of 0 %, 5 %, 10 %, 15 %, and 20 %.

To avoid ambiguity, it should be noted that these powers are expressed in terms of absolute, not relative, change. Thus, for example, they show the power of detecting an increase of 10 percentage points in the incidence rate as a function of background rate. For a background rate of 0, this is an increase from 0 % to 10 % incidence. For a background rate of 5 %, this is an increase from 5 % to 15 %. Of course, these changes could be expressed in terms of a decrease in the rate of non-occurrence, e.g. as decreases in survival rate. Simple arithmetic allows the use of these plots to compute relative rates of change as well.

In designing an experiment, no drastic change in number of replicates is generally needed to achieve at least 75 % power of observing the size effect deemed important, bearing in mind that we do not know in advance with certainty at which concentration this is evident from the power curves that if we are to estimate small changes in percent effects, then large sample sizes are needed.

In all cases shown, the test used is a 0,05-level test, as described elsewhere in this Technical Specification.

E.2.3 Example of power computations

An example of power computations using these plots might be helpful.

Suppose one wants to determine the NOEC for mortality in an experiment with *daphnia magna*, where past experience with this species and suggests the background mortality rate is near zero. The goal is to be able to detect a 20 % mortality rate. Suppose that based on preliminary range finding experiments, a decision was made to do an experiment with five concentrations at 50 ppm, 100 ppm, 200 ppm, 400 ppm and 800 ppm of the test compound plus a single (non-solvent) control. Furthermore, there is no anticipation of extra-binomial variance or within-tank correlations, so a standard Cochran-Armitage test can be done treating all subjects within a concentration equally (i.e. ignoring any tank or replicate effect). The question is then: how many subjects per concentration should be planned?

First, consider designs with the same number, n, of subjects in each concentration as in the control. The power of the Cochran-Armitage test depends on the shape of the dose-response curve, which we do not know. Powers have been simulated for numerous shapes. Based on an examination of the various power plots, a reasonable choice for design purposes is the linear dose-response shape. In addition, the power depends on the threshold of toxicity. For design purposes, we assume that is zero. The following plots help.

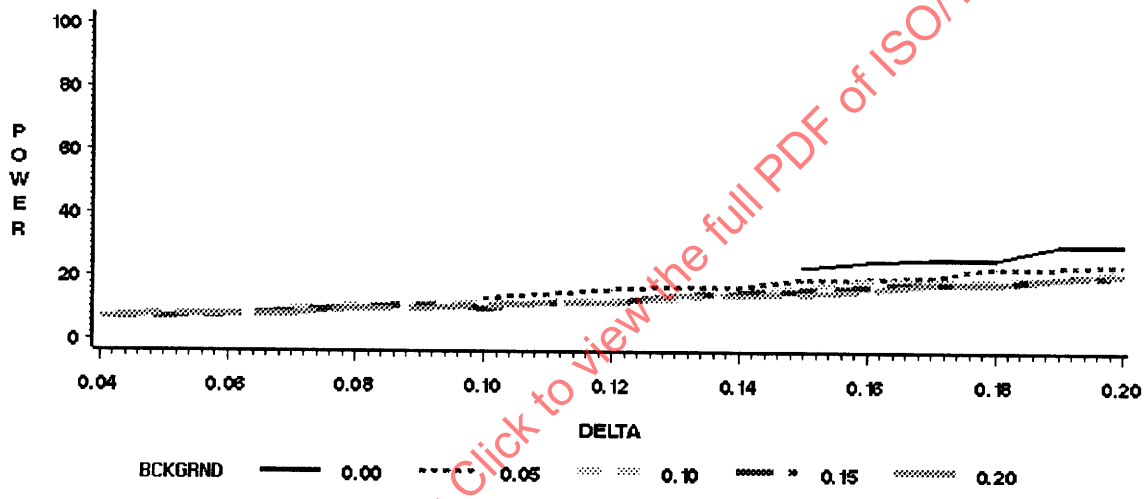


Figure E.1 — Cochran-Armitage test: Plot showing that 5 subjects per concentration would give very low power

Figure E.1 shows that 5 subjects per concentration would give very low power (about 25 %) to detect a 20 % change in the high concentration. There is little point in conducting the experiment for the purpose.

Consider a design with 20 subjects per concentration (Figure E.2).

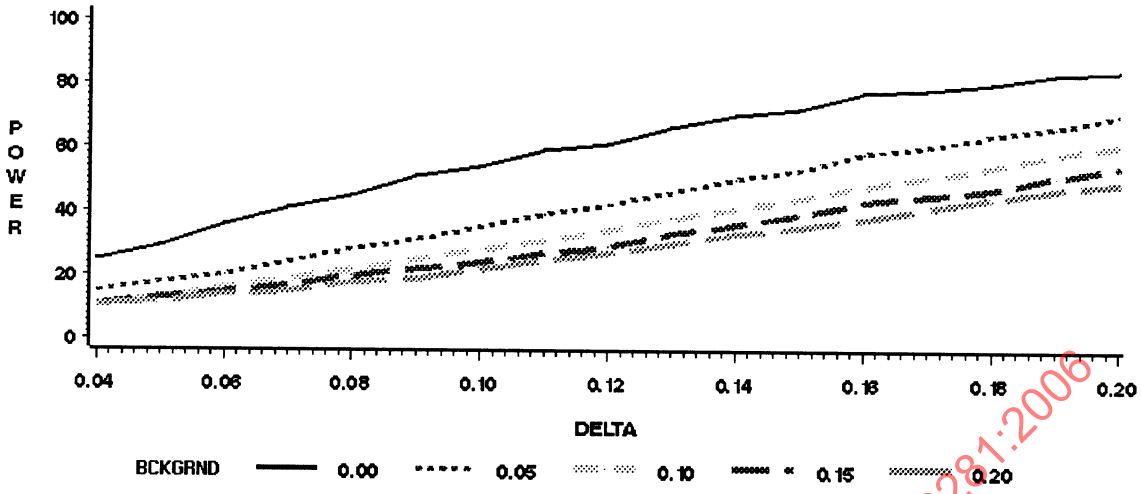


Figure E.2 — Cochran-Armitage test: Design with 20 subjects per concentration

This sample size gives a power of 82 % to detect a 20 % mortality in the 800 ppm concentration. This may well be adequate. Next, consider the power to detect a 20 % in lower concentrations. Fortunately, not much power is lost in the step-down procedure. The power to detect a 20 % mortality rate at 400 ppm is 80 %, at 200 ppm is 78 % and at 100 ppm is 76 %. However, if the background incidence rate were 10 % instead of zero, then the power to detect an increase in mortality rate of 20 % drops to around 40 %, which would be inadequate for most purposes.

E.2.4 Replicates

Decisions on the number of subjects per subgroup and number of subgroups per group should be based on power calculations using historical control data to estimate the relative magnitude of within- and among-subgroup variation and correlation. If there are no subgroups, then there is no way to distinguish housing effects from concentration effects and neither between- and within-group variances or nor correlations can be estimated, nor is it possible to apply any of the statistical tests described for continuous responses to subgroup means. Thus, a minimum of two subgroups per concentration is recommended; three subgroups are much better than two; four subgroups are better than three. The improvement in modelling falls off substantially as the number of subgroups increases beyond four. (This can be understood on the following grounds. The modelling is improved if we get better estimates of both among- and within-subgroup variances. The quality of a variance estimate improves as the number of observations on which it is based increases. Either sample variance has, at least approximately, a Chi-squared distribution. The quality of a variance estimate can be measured by the width of its confidence interval and a look at a Chi-squared table verifies the statements made.).

The number of subgroups per concentration and subjects per subgroup should be chosen to provide adequate power to detect an effect of magnitude judged important to detect. This power determination should be based on historical control data for the species and endpoint being studied.

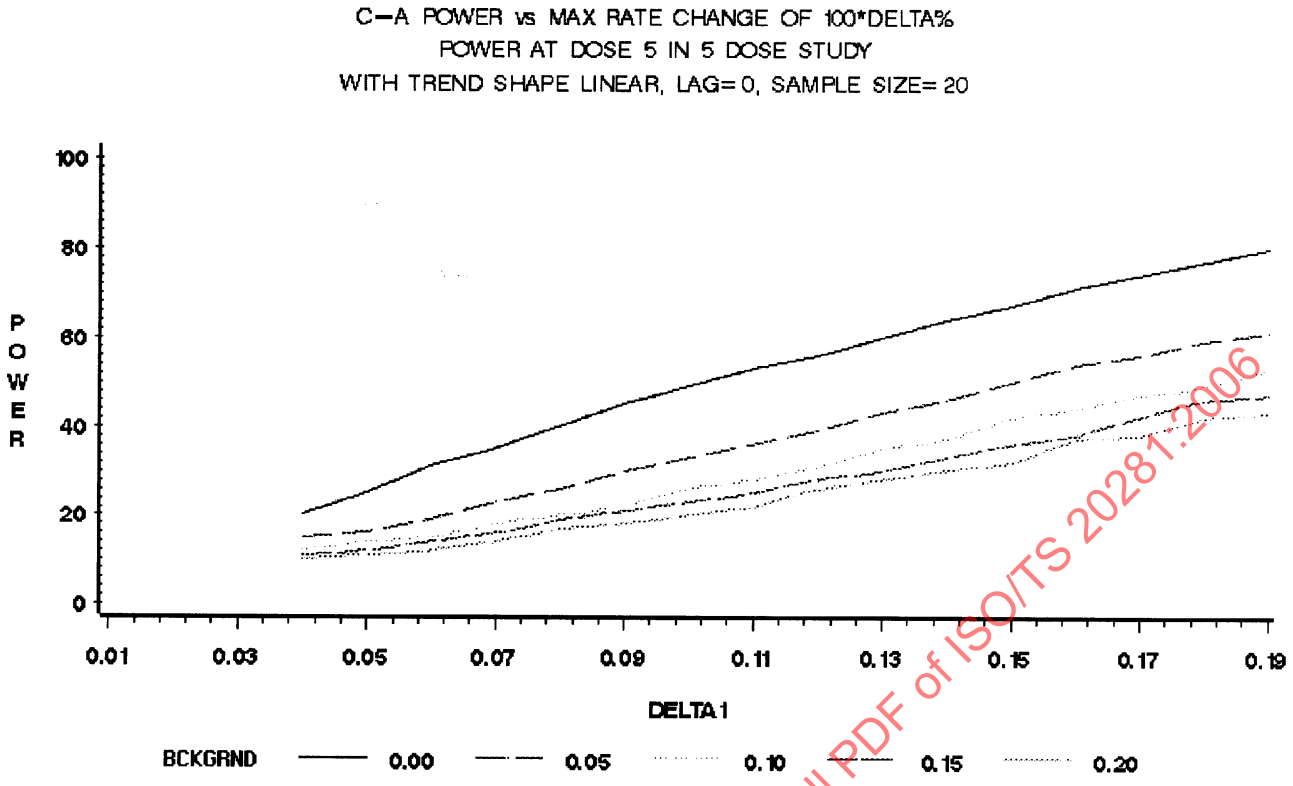


Figure E.3 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study, with trend shape linear, lag = 0, sample size = 20

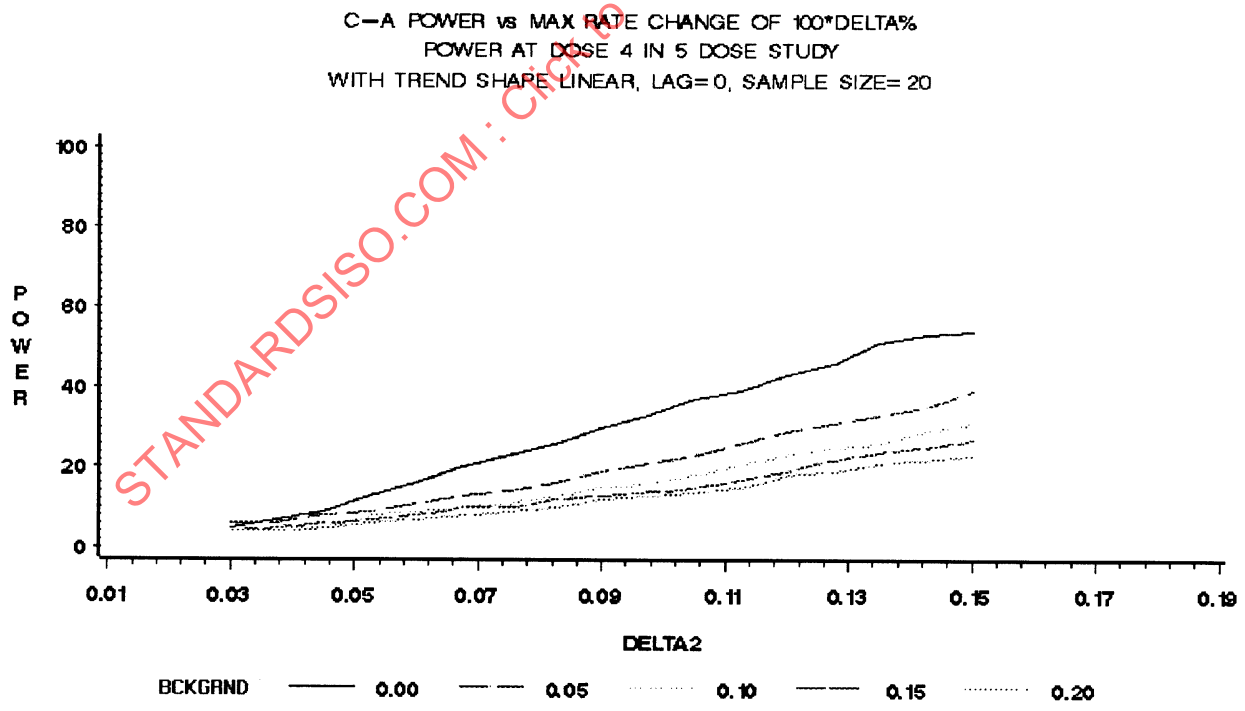


Figure E.4 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study, with trend shape linear, lag = 0, sample size = 20

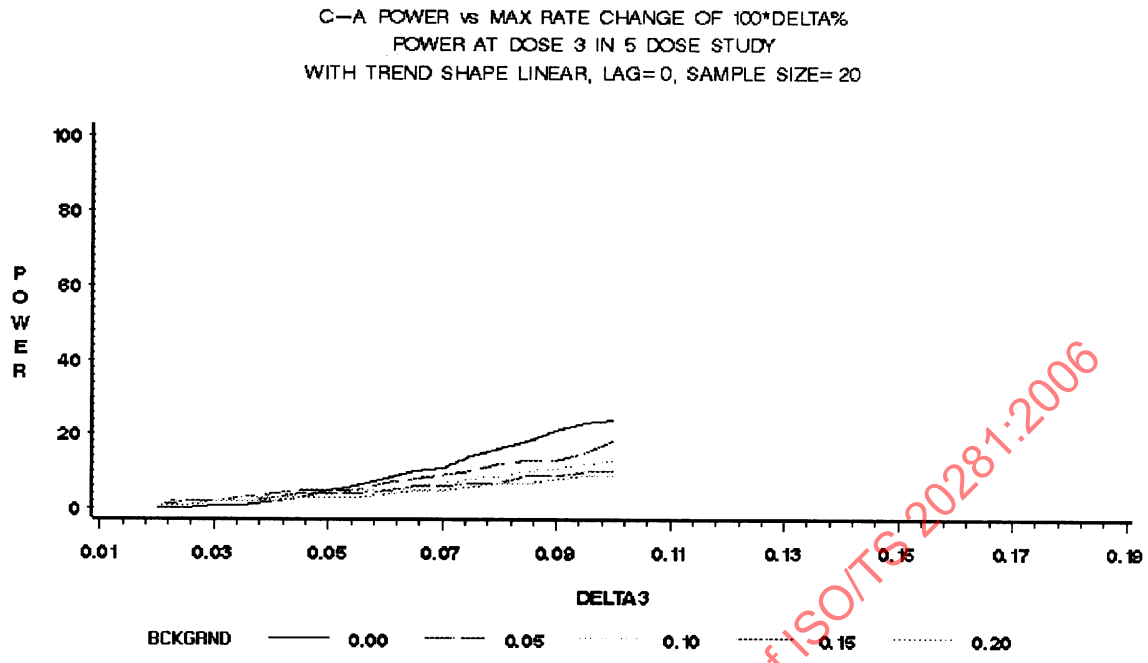


Figure E.5 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study, with trend shape linear, lag = 0, sample size = 20

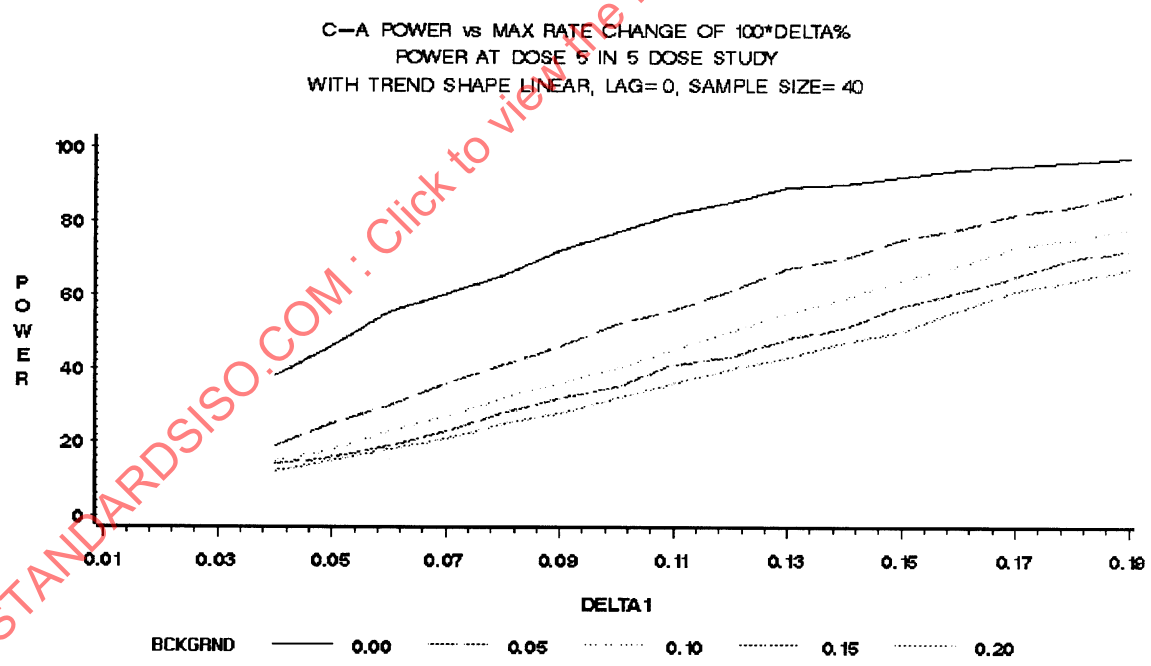


Figure E.6 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study with trend shape linear, lag = 0, sample size = 40

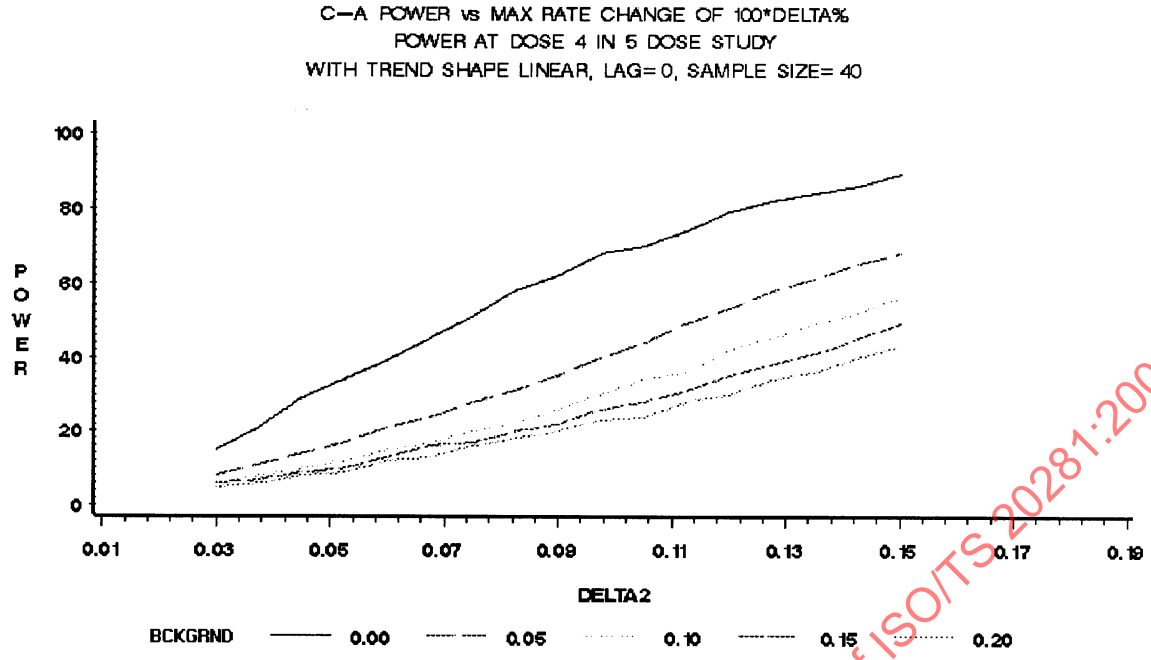


Figure E.7 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study with trend shape linear, lag = 0, sample size = 40

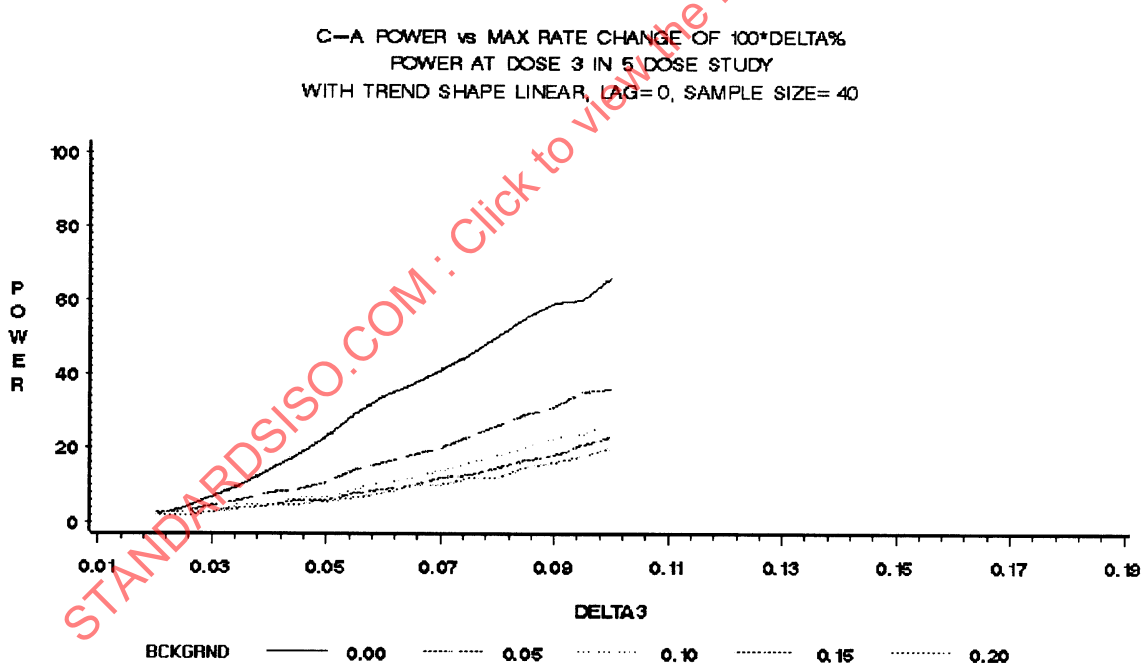


Figure E.8 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study with trend shape linear, lag = 0, sample size = 40

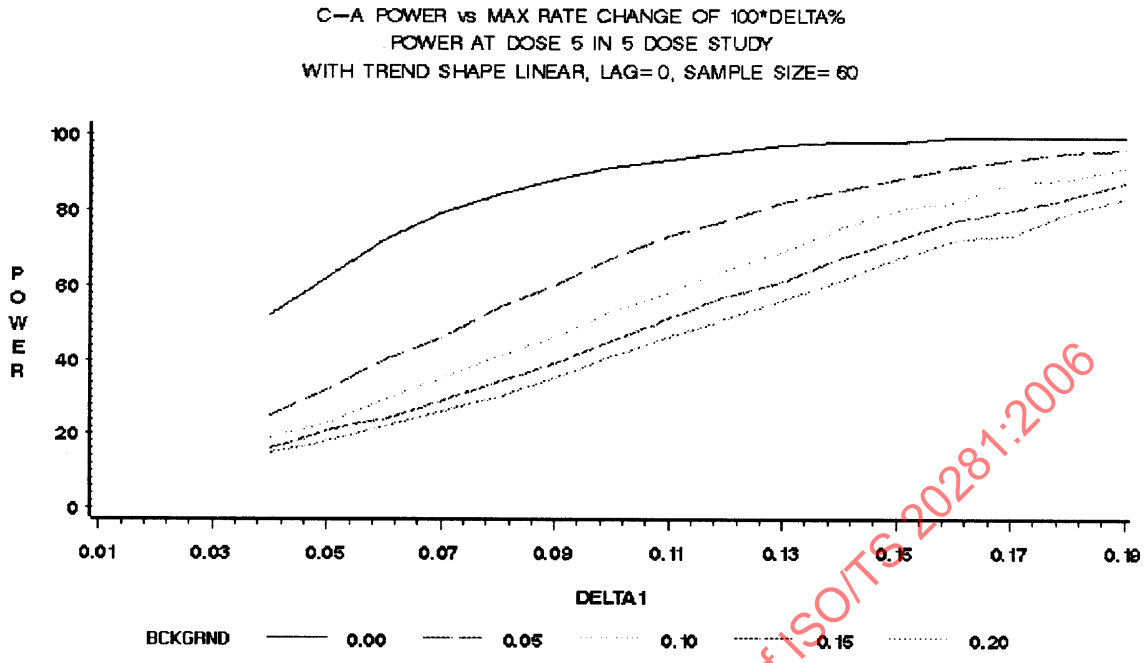


Figure E.9 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study with trend shape linear, lag = 0, sample size = 60

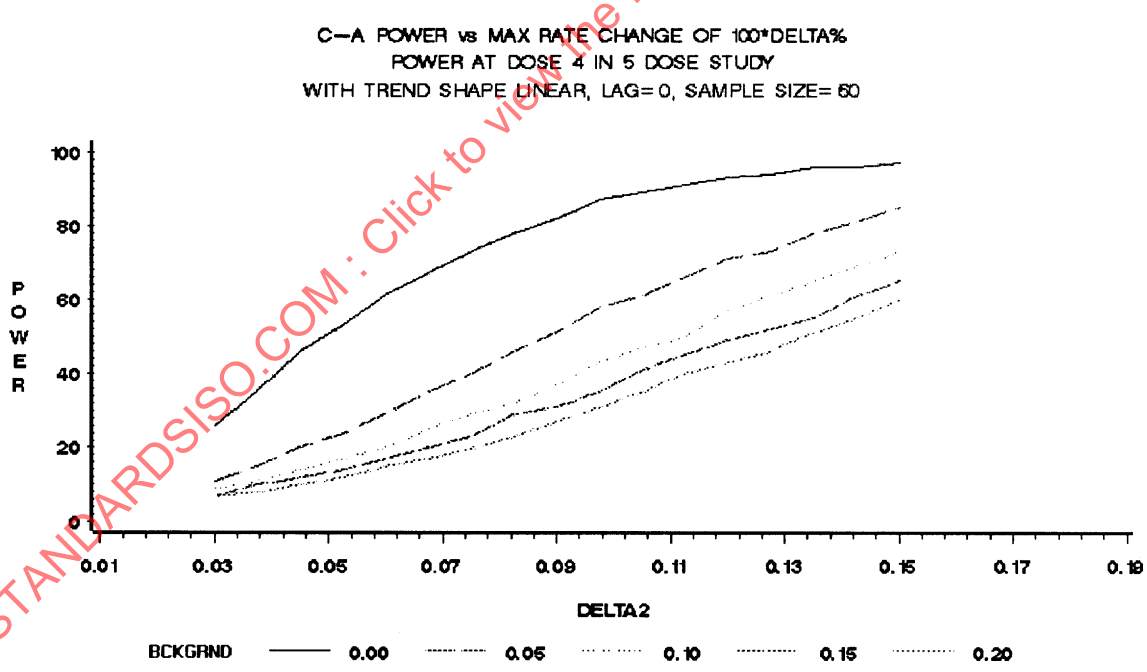


Figure E.10 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study with trend shape linear, lag = 0, sample size = 60

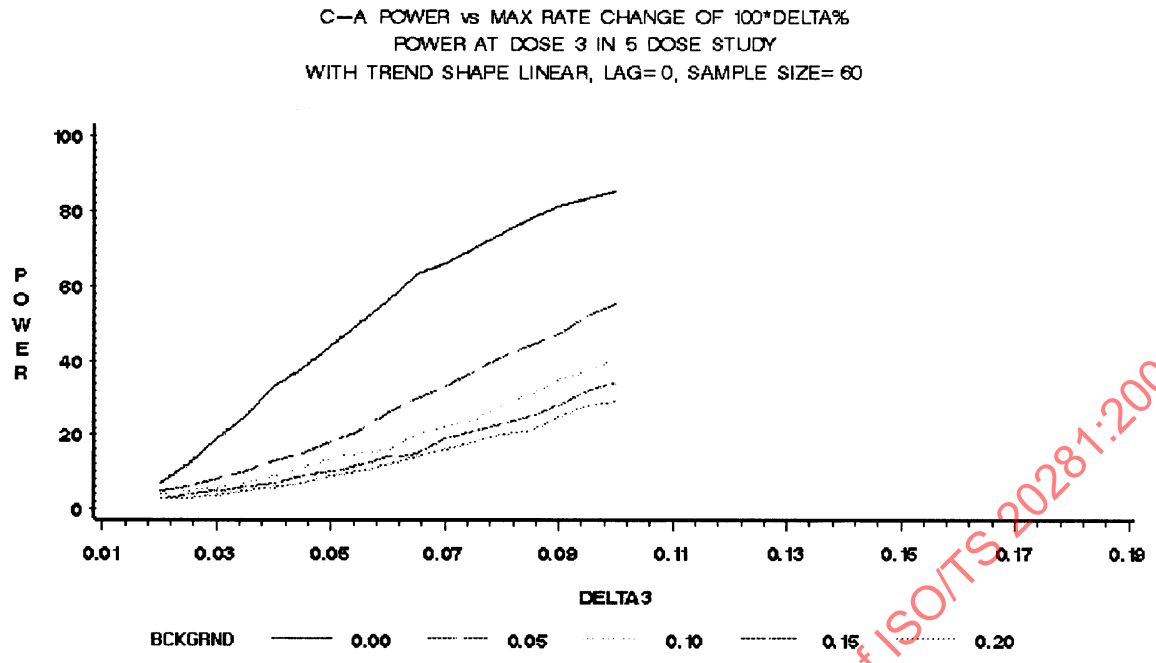


Figure E.11 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study with trend shape linear, lag = 0, sample size = 60

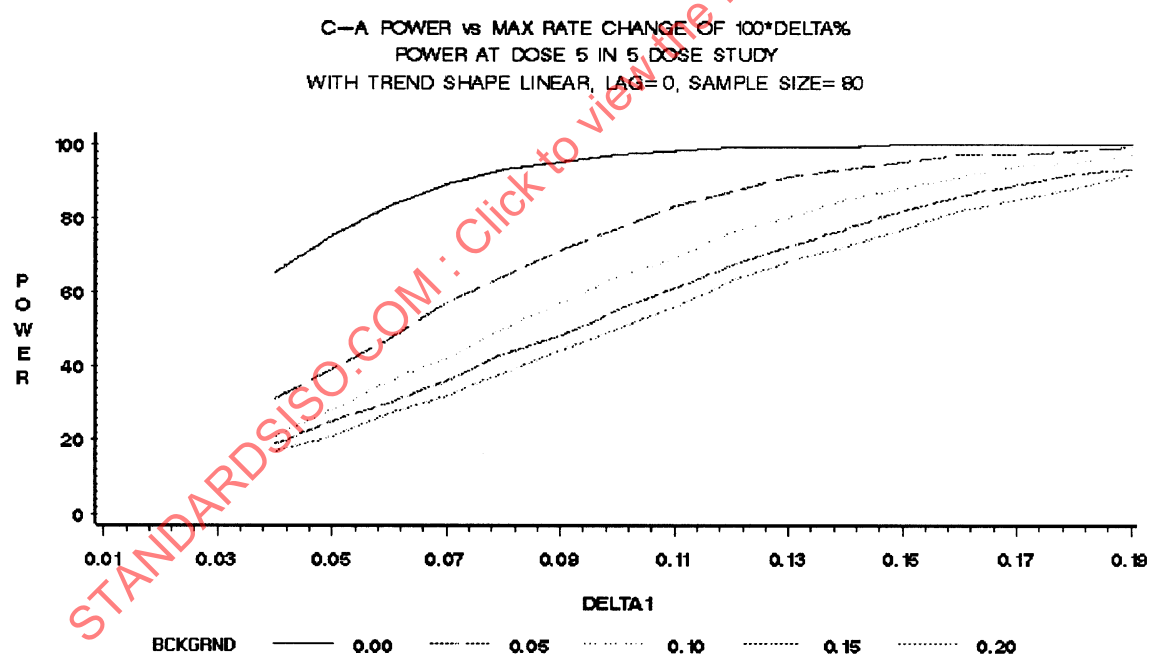


Figure E.12 — Cochran-Armitage power versus maximum rate change: Power at Dose 5 in 5-dose study with trend shape linear, lag = 0, sample size = 80

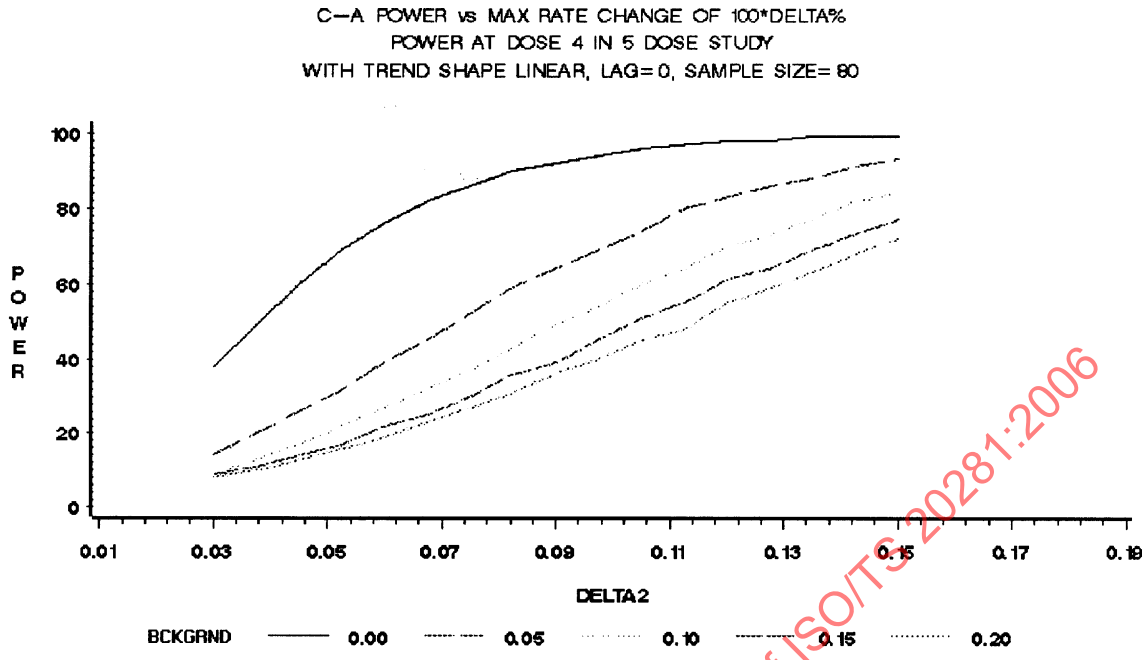


Figure E.13 — Cochran-Armitage power versus maximum rate change: Power at Dose 4 in 5-dose study with trend shape linear, lag = 0, sample size = 80

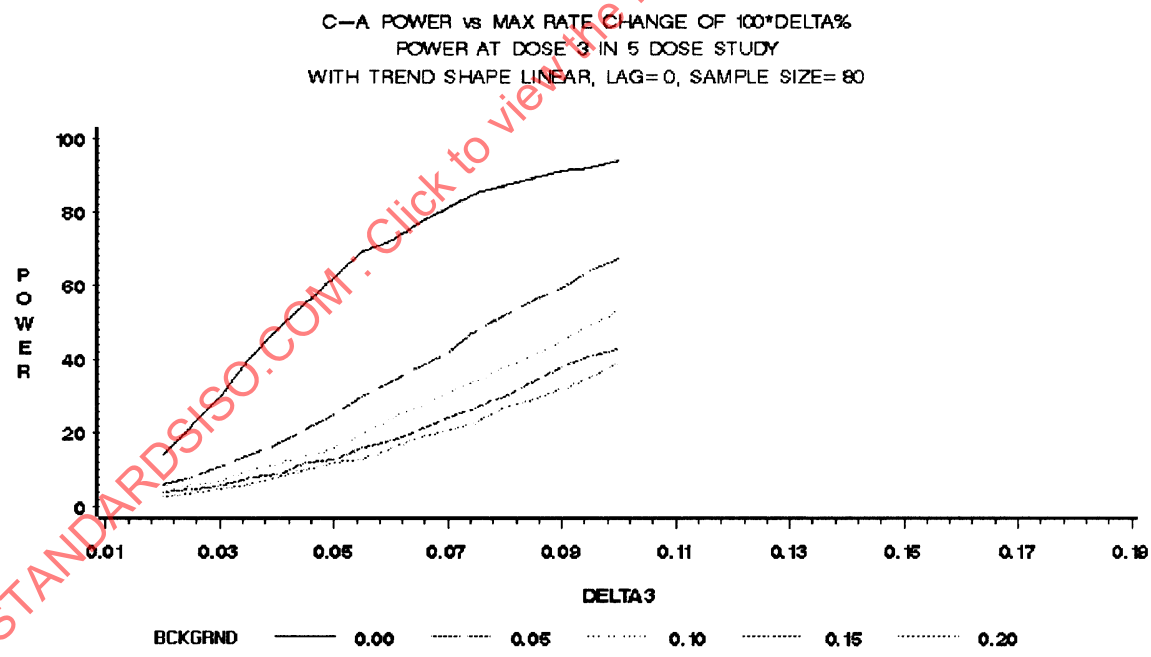


Figure E.14 — Cochran-Armitage power versus maximum rate change: Power at Dose 3 in 5-dose study with trend shape linear, lag = 0, sample size = 80

E.3 Description of selected tests for use with continuous data

E.3.1 References for selected tests

General references on trend tests are Barlow *et al.* (1972) and Robertson *et al.* (1988).

Hochberg and Tamhane (1987) discuss step-down tests in general.

Selwyn (1988) discusses the use of the Cochran-Armitage test in this way to establish the NOEC.

USEPA (1995) recommends the use of step-down trend tests to establish NOECs for both quantal (incidence) and continuous responses. USEPA (1995) recommends the Mantel-Haenszel test instead of the almost equivalent Cochran-Armitage test for incidence data and Jonckheere-Terpstra test for continuous data to establish the NOEC.

Dunnett and Tamhane (1991, 1992, 1995) discuss step-down trend tests to determine the equivalent of NOEC in medical tests.

Tamhane and Dunnett (1996) discuss them in toxicology experiments, as do Tamhane *et al*. (2001), Capizzi *et al*. (1984), Tukey *et al.* (1985). These authors criticize single-step procedures as having low power of detecting real effects and offer step-down procedures as an improvement.

Both 1- and 2- tailed step-down procedures belong to the class of “fixed sequence” tests in the terminology of Westfall (1999) and Westfall *et al*. (1999). Such tests also belong to the more general class of *closed systems of hypothesis tests* Peritz (1970) and Marcus, Peritz and Gabriel (1976). Budde and Bauer (1989) discuss a step-down procedure based on the Jonckheere-Terpstra test that differs from that discussed here. Kodell and Chen (1991) apply this same idea to quantal data, including the Cochran-Armitage test. These authors followed the more general but also more cumbersome closed system of Peritz (1970) and Marcus, Peritz and Gabriel (1976) rather than the fixed sequence approach.

There are several methods used to perform the tests of “fixed sequence” hypotheses for continuous responses, including Williams’ test, the Jonckheere-Terpstra test, Bartholomew’s test, Welch, and Brown-Forsythe tests, and sequences of linear contrasts, among others. These are well established as tests of the stated hypothesis in the statistics literature.

Williams’ test is also a step-down trend test commonly used to establish NOECs in toxicological experiments. Tamhane and Dunnett (1996) discuss Williams’ test and two new step-down procedures for use in toxicology and drug development. They note the low power of multiple comparison approaches to this work in toxicology in the context of discussing the benchmark dose approach of Gaylor (1983) and Crump (1984). They claim Williams’ test loses power under some non-monotone alternatives, while their methods do not.

Salsburg (1986) discusses the low power of ANOVA methods in analysing dose-response experiments. He recommends Bartholomew’s test as the most general test against ordered alternatives. He discusses the use of linear contrasts, but notes that they may not be powerful in experiments where the lower doses have no effect and all the effect is found only at the highest dose. Contrast tests make no direct use of the supposed monotone dose-response relationship, and, hence, are lower in power than alternative procedures that do. Also, linear contrasts test specifically for a linear relationship, not monotone relationships. If the dose-response is not linear with respect to the particular dose metric used, it loses power. Bartholomew’s test is difficult to implement and Puri (1965) has shown that Jonckheere-Terpstra test is of very similar power. Robertson *et al*. (1988) has shown that under some conditions, Jonckheere-Terpstra test is more powerful, does not assume any specific shape (such as linearity) in the dose-response, and only requires monotonicity. It is also easily incorporated into step-down procedures. Robertson *et al*. (1988) discuss in great depth various tests that are appropriate in dose-response experiments, or more generally, when the explanatory variable has an order restriction. They found that under certain conditions, namely the mean responses are approximately uniformly related to dose order (not magnitude); Jonckheere-Terpstra is more powerful than other alternatives considered.

Bartholomew (1961) compares Jonckheere-Terpstra test to his own. In the case of 3 or 4 treatment groups, Bartholomew sees little difference between the power of the two tests, for either equally spaced means or all but one mean equal. He, in fact, found Jonckheere-Terpstra test to be preferable, given its distribution-free nature. For larger k, however, Bartholomew's test is superior for the case of all means but one equal, while Jonckheere-Terpstra is still preferable for the equally spaced means case.

E.3.2 Williams' Test

E.3.2.1 General

Williams' test is step-down or fixed-sequence test procedure that can be used in the same situations as the Jonckheere-Terpstra test. Unlike the latter, Williams' is based on normally distributed, homogeneous responses and formally incorporates the presumed monotone dose-response in the estimated mean effects at each dose. These means are called isotonic estimates and are based on maximum likelihood theory, given the dose-response is monotone. Isotonic estimators were developed by Ayer *et al.* (1955), who called their method Pool-the-Adjacent-Violators (PAVA) algorithm. Isotonic regression was introduced by Barlow *et al.* (1972).

E.3.2.2 Assumptions

Independent random samples of normally distributed, homogeneous variables with monotone means (for example, $\mu_0 \leqslant \mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_k$). The maximum likelihood estimate of μ_i under the monotone assumption is given by

$$\tilde{\mu}_i = Max_{1 \leqslant u \leqslant i} \; Min_{i \leqslant v \leqslant k} \frac{\sum_{j=u}^{v} n_j \bar{Y}_j}{\sum_{j=u}^{v} n_j},$$

where

$\bar{Y}_j$ is the arithmetic mean of dose group i; and

n_j is the sample size.

The roles of Min and Max are reversed for a non-increasing trend. It is easier to compute isotonic estimators than to describe them. Given that we expect a non-decreasing trend in the means, we look for a violation of this expected result. If $\bar{Y}_j > \bar{Y}_{j+1}$, then we pool (or amalgamate) these two means using a simple weighted average. We then re-examine the reduced set of means for violations of the expected order and do additional pooling as needed. It makes no difference in the final result which adjacent means we amalgamate first. We continue this amalgamation procedure until the means are in the expected non-decreasing order. The control is never amalgamated with positive dose groups.

It is evident from this description that, given unequal sample sizes, it can happen that doses i and i+1 are amalgamated and $\bar{t}_{i+1}$ is significantly different from the control but $\bar{t}_i$ is not. There are several ways to avoid this situation. Williams suggests re-computing the isotonic means at each stage of the step-down procedure, using only those means remaining at each stage. In the balanced case, this is equivalent to the procedure already described. Another alternative is to use only the reduced set of amalgamated means and declaring all dose groups involved in the amalgamated mean to have significant effects if the amalgamated mean is significantly different from the control. Williams also suggests other modifications.

E.3.2.3 Advantages of Williams' test

This test makes direct use of the assumed monotone dose-response both in terms of the estimated mean effect at each dose and in the step-down conduct of the test. Under one of the modifications indicated above, it cannot happen that a low dose shows an effect and a higher dose does not. The test can be modified to take multiple sources of variation into account. The same method used in the Tamhane-Dunnett test below can be used for this purpose.

E.3.2.4 Disadvantages of Williams' test

This test loses power when additional higher dose groups are added to the design, unless the effects at the new doses are substantially greater than at the lower dose levels. There can be a loss in power if there is a change in direction at the high dose, such as might occur if there is substantial mortality in that group. It is affected in an unknown way if the data are not normally distributed or heterogeneous. (However, Shirley's (1979) non-parametric alternative to Williams' is available for non-normal or heterogeneous data.) According to Bretz (1999) and Bretz and Hothorn (2000), the null distribution of Williams' test is known only for the balanced case. For the unbalanced case, the actual p-value when the nominal is 0,05 can be as much as 20 % larger than the nominal. Williams (1972) claims that the equal sample size error probabilities are approximately correct if the difference in sample sizes is not great. Bretz and Hothorn (2000) give reasons why this test should not be used for highly unbalanced data and they provide an alternative test, similar to Williams but overcoming several difficulties.

E.3.2.5 Power of Williams' test:

Definitive power properties of Williams' test are not readily available, though limited simulations have been published [Marcus, R. (1976); Poon, A.H. (1980); Shirley, E.A. (1979); Williams, D.A. (1971, 1972)] that suggest power characteristics similar to those for the Jonckheere-Terpstra test, so long as the data meet the requirements for Williams' test and there is no change in direction at the high dose. Large-sample theoretical power properties have been published by Puri (1965), Bretz (1999) and Bretz and Hothorn (2000).

E.3.2.6 Confidence intervals for NOEC from Williams' test

It is sometimes of interest to have confidence bands for a dose-response curve. These bands can be used to construct simultaneous confidence intervals for mean or individual responses at different doses, form confidence intervals for doses at a given response, and to compare different dose-response curves. In the parametric regression context, it is quite straight forward to construct such confidence bands. When no such model is fit to the data, other methods shall be employed.

Genz and Bretz (1999) have shown how simultaneous confidence intervals can be computed for the mean effect at each concentration that are applicable to Williams' and Dunnett's tests, as well as to various others. These procedures may suffer from the need to compute confidence intervals for means of no interest, namely at all concentrations rather than just at the NOEC. Hence, these intervals are wider than what might be obtained by focusing on just the NOEC. The methods of Korn, Williams, and Schoenfeld which are described below are appropriate in the context of Williams' test.

According to Korn (1982), for a normally distributed response, one could construct simultaneous 1-α confidence intervals for the means μ_i from the following.

$$\bar{Y}_j - m_{k,N-k}\, s / \sqrt{n_j} < \mu_j < \bar{Y}_j + m_{k,N-k}\, s / \sqrt{n_j} \qquad \text{(E.1)}$$

where

$m_{k,N-k}$ is the studentized maximum modulus distribution (Hochberg and Tamhane, 1987) for k treatment groups and $N–k$ degrees of freedom;

s is the square-root of the pooled within-group variance estimate;