Unclassified **ENV/MC/CHEM(98)18**

Organisation de Coopération et de Développement Economiques
Organisation for Economic Co-operation and Development

OLIS : 27-Jan-1998
Dist. : 28-Jan-1998

Or. Eng.

ENVIRONMENT DIRECTORATE
CHEMICALS GROUP AND MANAGEMENT COMMITTEE

OECD SERIES ON TESTING AND ASSESSMENT
Number 10

Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data

61088

Document complet disponible sur OLIS dans son format d'origine
Complete document available on OLIS in its original format

OECD Environmental Health and Safety Publications

Series on Testing and Assessment

No. 10

-

Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data

Environment Directorate

Organisation for Economic Co-operation and Development

Paris 1998

Also published in the Series on Testing and Assessment:

No. 1, *Guidance Document for the Development of OECD Guidelines for Testing of Chemicals* (1993; reformatted 1995)

No. 2, *Detailed Review Paper on Biodegradability Testing* (1995)

No. 3, *Guidance Document for Aquatic Effects Assessment* (1995)

No. 4, *Report of the OECD Workshop on Environmental Hazard/Risk Assessment* (1995)

No. 5, *Report of the SETAC/OECD Workshop on Avian Toxicity Testing* (1996)

No. 6, Report of the Final Ring Test of the Daphnia magna *Reproduction Test* (1997*)*

No. 7, Guidance Document on Direct Phototransformation of Chemicals in Water (1997)

No. 8, Report of the OECD Workshop on Sharing Information about New Industrial Chemicals Assessment (1997)

No. 9, Guidance Document for the Conduct of Studies of Occupational Exposure to Pesticides During Agricultural Application (1997)

No. 11, Detailed Review Paper on Aquatic Testing Methods for Pesticides and Industrial Chemicals.
Part I: Report. Part II: Annexes (1997)

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 29 industrialised countries in North America, Europe and the Pacific, as well as the European Commission, meet to co-ordinate and harmonize policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialized Committees and subsidiary groups composed of Member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's Workshops and other meetings. Committees and subsidiary groups are served by the OECD Secretariat, located in Paris, France, which is organised into Directorates and Divisions.

The work of the OECD related to chemical safety is carried out in the Environmental Health and Safety Programme. As part of its work on chemical testing, the OECD has issued several Council Decisions and Recommendations (the former legally binding on Member countries), as well as numerous Guidance Documents and technical reports. The best known of these publications, the **OECD Test Guidelines**, are a collection of methods used to assess the hazards of chemicals and of chemical preparations such as pesticides and pharmaceuticals. They cover tests for physical and chemical properties, effects on human health and wildlife, and accumulation and degradation in the environment. The OECD Test Guidelines are recognised worldwide as the standard reference tool for chemical testing.

More information about the Environmental Health and Safety Programme and its publications is available on the OECD’s World Wide Web site (see next page).

The Environmental Health and Safety Programme co-operates closely with other international organisations. This document was produced within the framework of the Inter-Organization Programme for the Sound Management of Chemicals (IOMC).

The Inter-Organization Programme for the Sound Management of Chemicals (IOMC) was established in 1995 by UNEP, ILO, FAO, WHO, UNIDO and the OECD (the Participating Organizations), following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. UNITAR joined the IOMC in 1997 to become the seventh Participating Organization. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organizations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

For the complete text of this and many other Environmental Health and Safety publications, consult the OECD's World Wide Web site (http://www.oecd.org/ehs/)

or contact:

OECD Environment Directorate,
Environmental Health and Safety Division

2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 45 24 16 75

E-mail: ehscont@oecd.org

FOREWORD

This document contains the report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data which took place in Braunschweig, Germany, in October 1996.

The Joint Meeting of the Chemicals Group and the Management Committee of the Special Programme on the Control of Chemicals recommended that this document be derestricted. It is being published on the responsibility of the Secretary-General of the OECD.

TABLE OF CONTENTS

(continued on next page)

EXECUTIVE SUMMARY

Following a decision taken by the National Co-ordinators of the OECD Test Guideline Programme and the OECD Risk Assessment Advisory Body (RAAB) at their joint session in December 1995, an OECD Workshop on Statistical Analysis of Aquatic Ecotoxicity Data was held in Braunschweig, Germany on 15-17 October 1996. The workshop was hosted by the *Biologische Bundesanstalt für Land- und Forstwirtschaft (*BBA) in Braunschweig and was chaired by Dr Arno Lange of the German *Umweltbundesamt* (UBA).

The objectives of the workshop were:

- to review the options available for the analysis of data from ecotoxicity tests;
- to compare their advantages and disadvantages;
- to recommend (a) the most appropriate approach for deriving a summary parameter(s) which has (have) scientific validity, and (b) further work to be undertaken by the OECD and/or others, as appropriate.

In plenary and working group sessions participants discussed statistical data analysis appropriate for single-species chronic/subchronic studies using a number of test concentrations, i.e. dose-response tests. Aquatic tests served as a basis for these discussions, however the issues addressed may be similar for toxicity tests in general. Background documents had been prepared on the following main existing data analysis approaches for such tests: analysis of variance/hypothesis testing ("ANOVA/NOEC approach");[1] regression analysis (based on empirical models); and mechanistic modelling (theory-based).

It was concluded that the NOEC, as the main summary parameter of aquatic ecotoxicity tests, is inappropriate for a number of reasons (see detailed discussion in the report) and should therefore be phased out. It was recommended that the OECD should move towards a regression-based estimation procedure. The time course of effects should be incorporated in the analytical procedures, and the OECD should initiate a study of the available time-dependent regression models (both mechanistic and empirical) in order to select those which best meet its needs. The study should also address the issue of appropriate values of x for EC_x and the optimal experimental designs. A steering group should be set up to direct the mathematical, statistical and biological work required to take the workshop recommendations forward. This group should include representatives from the appropriate scientific and regulatory communities.

[1] For definitions, see the Glossary.

RESUME

Suite à une décision des Coordinateurs nationaux du Programme de l'OCDE sur les Lignes directrices pour les essais et du Groupe consultatif de l'OCDE sur l'évaluation des risques (RAAB) lors de leur session conjointe en décembre 1995, un atelier de l'OCDE sur l'analyse statistique des données d'écotoxicité aquatique s'est tenu à Braunschweig, en Allemagne du 15 au 17 octobre 1996. L'atelier fut accueilli par le *Biologische Bundesanstalt für Land-und Forstwirtschaft* (BBA), sous la présidence du Docteur Arno Lange du *Umweltbundesamt* (UBA). L'atelier avait pour objectifs de:

- examiner les différentes options existantes pour l'analyse des données des essais d'écotoxicité;
- comparer leurs avantages et leurs inconvénients;
- faire des recommandations concernant (a) l'approche la plus appropriée pour obtenir un/des paramètre(s) "résumé(s)" qui soi(en)t scientifiquement valides, et (b) les travaux ultérieurs qu'il conviendrait que l'OCDE et/ou d'autres mettent en oeuvre;

Au cours des sessions plénières et des sessions de groupe de travail, les participants ont débattu de l'analyse statistique des données relatives aux études chroniques/subchroniques réalisées sur un seul organisme et utilisant un certain nombre de concentrations, c'est-à-dire les essais dits "dose-réponse". Les essais dans le domaine aquatique ont servi de base aux discussions; cependant, les questions abordées peuvent être les mêmes pour les essais de toxicité et concernent les principales approches pour l'analyse des données d'essais que sont: l'analyse de la variance/hypothèse d'essai ("approche ANOVA/NOEC")[2], l'analyse de régression (basée sur les modèles empiriques), et la modélisation méchanistique (basée sur la théorie).

Il a été conclu que, pour un certain nombre de raisons (voir les arguments détaillés dans ce rapport), la NOEC ne convient pas comme principal paramètre "résumé" des essais d'écotoxicité aquatique et qu'elle devrait donc être progressivement abandonnée. Il a été recommandé que l'OCDE s'oriente vers une procédure d'estimation basée sur la régression. Les effets au cours de temps devraient être pris en compte dans les procédures analytiques, et l'OCDE devrait mettre en place une étude des modèles disponibles de régression dépendants du temps (à la fois méchanistique et empirique) afin de choisir ceux qui correspondent le mieux aux besoins de l'OCDE. L'étude devrait aussi aborder la question des valeurs appropriées de *x* pour l'EC_x et celle de la conception optimale des essais. Un groupe directeur devrait diriger les travaux d'ordre mathématique, statistique et biologique nécessaires afin de mettre en place les recommandations de l'atelier. Ce groupe devrait être composé de représentants des communautés scientifiques et réglementaires compétentes.

[2] Se reporter au glossaire pour les définitions

INTRODUCTION

Background

Within OECD countries, a number of aquatic ecotoxicity test guidelines are used to assess the potential effects of chemicals (including pesticides) on aquatic organisms. For use in hazard/risk assessment schemes, summary parameters (e.g. the LC_{50} and the NOEC) are established by statistical methods. Statistical evaluation plays a major role in developing test guidelines, since the experimental design is crucial for the statistical method that can be applied, and both together are central for developing tests that produce high-quality data with a minimum use of resources and test organisms.

In 1992, the National Co-ordinators of the OECD Test Guideline Programme decided that existing and draft aquatic toxicity test guidelines with respect to test design and statistical data analysis should be reviewed. “A Review of Statistical Data Analysis and Experimental Design in OECD Aquatic Toxicology Test Guidelines” (included as an annex to this report), prepared by Dr Simon Pack in 1993, was widely circulated and discussed. The review and the recommendations made were broadly well received and appreciated. Other activities of the scientific community in regard to this issue included workshops in The Hague (1994) and London (1996).[3]

In all of these activities it was concluded that the NOEC is inappropriate as a summary measure of toxicity. Replacing the NOEC as suggested has implications for test designs as well as for hazard/risk assessment. Hence, in the context of OECD work, both the OECD Test Guideline and Risk Assessment Programmes became involved and the relevant bodies (the National Co-ordinators of the Test Guideline Programme and the Risk Assessment Advisory Body, or RAAB) decided there should be a joint workshop to investigate the issue further.

The OECD Workshop on Statistical Analysis of Aquatic Ecotoxicity Data was held in Braunschweig, Germany on 15-17 October 1996. It was hosted by the BBA and was chaired by Dr Arno Lange from the German UBA. There were over 50 participants representing the governments of 15 Member countries and industry (BIAC, ECETOC, GCPF/ECPA) [see the participants’ list included as an annex 1].

Objectives

The objectives of the workshop were: to review the options available for the analysis of data from ecotoxicity tests; to compare their advantages and disadvantages; and to recommend (a) the most appropriate approach for deriving a summary parameter(s) which has (have) scientific validity, and (b) further work to be undertaken by the OECD and/or others, as appropriate.

[3] F. Noppert, N. Van der Hoeven and A. Leopold, How to measure no effect: towards a new measure of chronic toxicity in ecotoxicology. Workshop Report of the Netherlands Working Group on Ecotoxicology, The Hague, 1994; P.F. Chapman, M. Crane, J.A. Wiles, F. Noppert and E.C. McIndoe (eds.), Asking the Right Questions: Ecotoxicology and Statistics. Report of a Workshop Held at Royal Holloway University of London, Surrey, UK, SETAC-Europe, 1996.

Focus

The workshop focused on approaches to data analysis appropriate for single-species chronic/subchronic *aquatic tests* using a number of test concentrations, although it was recognized that the discussion might also be relevant to the analysis of data from ecotoxicity tests in general. The implications of the different statistical approaches for test design were also considered.

Workshop structure and discussion topics

The workshop was organised around a series of plenary sessions and three working groups (see the table showing the working groups' composition, which has been included as an annex). The working groups each addressed the same issues, in parallel sessions, and then reported on their progress during plenary sessions. The discussion topics were framed as questions:

Session 1:

- Why are OECD ecotoxicity tests performed?
- What kind of information do we want from the tests?
- Are we happy with the current statistical practices?

Session 2:

- Review and comparison of the different approaches to data analysis:
  - ◊ Should the NOEC be retained?
  - ◊ Which other analytical technique could replace the NOEC?
  - ◊ What type of information (statistical summary parameter; test endpoint) do we want from the new approach?
- What work needs to be done with respect to the selection of statistical approach?

The reports of each of the three working groups have been included in the form of annexes.

Background documents

The following documents describing the three main approaches to data analysis (including an assessment of their strengths and weakness) were prepared and distributed in advance of the workshop:

- "A Discussion of the NOEC/ANOVA Approach to Data Analysis" by Simon Pack;
- "Alternatives to the NOEC Based on Regression Analysis" by Peter Chapman;
- "Dynamic Measures for Ecotoxicity" by S.A.L.M. Kooijman, et al.;

- "The Dynamic Energy Budget (DEB) Model" by S.A.L.M. Kooijman.

Each of these documents is included in this report as an annex. "A Review of Statistical Data Analysis and Experimental Design in OECD Aquatic Toxicology Test Guidelines" by Simon Pack (1993), which has until now remained in the "gray" literature, is also included as an annex.

SUMMARY OF PLENARY DISCUSSIONS

Plenary Rapporteurs: Peter Chapman and Mark Crane (United Kingdom)

This section summarises the main outcomes of the workshop. The discussion issues, together with a summary of the views put forward and a list of recommendations arising out of the workshop, are listed below. There was a high degree of agreement on all of the issues. Where there was disagreement or lack of consensus, this is highlighted. A more detailed account of the proceedings of each working group can be found in their individual reports, which are included as annexes.

1. Why are OECD ecotoxicity tests performed?

Ecotoxicity tests are performed to evaluate the toxicity of chemicals, in order to predict their potential effects on natural populations. These tests provide information which is used in the registration/notification of new chemicals and the assessment of older chemicals, including pesticides and biocides. The results of these tests are used in the classification and labelling of chemicals and contribute the "effects" component to a risk assessment. They may also be used to predict adverse effects in the event of an accident.

OECD Test Guidelines are primarily developed for the above reasons but may also be used in other situations, including the bioassay of environmental samples and fundamental research.

2. What kind of information do we need from OECD ecotoxicity tests?

The workshop agreed that toxicity tests should provide information that is accurate and precise, and that permits easy interpretation by the non-expert. Ideally, information on the time course of effects should be integrated with information on the concentration-response curve. The information should be of "biological relevance", although there was no consensus on whether this should refer simply to the types of measurements taken (e.g. survival, growth and reproduction are usually considered as relevant parameters because changes in them can affect population abundance) or to "ecological relevance".

It was recognised that the specific type of information required from a test will largely depend on the way in which it is extrapolated to the natural world. There was no consensus on whether "classification" and "risk assessment" demanded different information and analyses, or whether classification is simply a point on the road to risk assessment which uses similar data and analyses. However, classification will normally use only acute lethal data, while risk assessment will often use both lethal and sublethal data.

3. Are we happy with current statistical practices?

There was virtually unanimous agreement that current practices were unsatisfactory, and there was a great deal of consistency in the views put forward. The list below is a comprehensive summary of all the views presented:

- There is a concern that the NOEC may not be sufficiently protective because of the danger of false negatives.

- The statistical methods are suboptimal.

- OECD Test Guidelines contain insufficient information on statistical techniques.

- Data are wasted in the determination of values such as the NOEC.

- NOECs are leading to misunderstandings and misinterpretations.

- Current summary statistics cannot be linked to population models.

- There are no statements of biological significance, only of statistical significance (there was no consensus on this issue, as “biological significance” appears to mean different things to different people).

- More effective use should be made of test animals.

- Results are often imprecise.

- Biologically relevant covariates are not considered.

4. What should OECD tests look like in future?

It was agreed that new testing frameworks should not be developed that (a) exclude results from tests performed within the current framework, or (b) remove all flexibility in approach. However, there was common agreement that certain improvements are desirable:

- The danger of false negatives should be reduced.

  Tests should focus on biologically significant endpoints (although a definition needs to be agreed upon first).

- We should be able to link test results to predictive ecological/biological models.

- Biologically important covariates should be included, where this is appropriate.

- Better use should be made of all test measurements.

- Test guidelines should make explicit recommendations on appropriate statistical analytical techniques.

- Summary statistics and parameters should be capable of being interpreted by non-statisticians and decision-makers.

5. Comparison of the different approaches to data analysis

5.1 Hypothesis testing to determine a NOEC

The NOEC is familiar and, to date, has been widely used as a basis for risk assessment. It is perceived to be easy to understand, and software is freely available. However, hypothesis testing in general is not well suited to the type of data obtained from most toxicity tests (with the possible exception of limit tests) and the NOEC, in particular, is statistically unfounded (the annexed review report by Simon Pack gives a detailed explanation of this point).

5.2 Static regression

Static regression models, in which a model is fitted to measurements taken at a single fixed time, enable EC_x values to be estimated for each time of assessment. However, because these models do not include a time component, they do not use all of the data in an efficient manner. For those tests which are assessed only once, they are the only option available.

5.3 Time-dependent regression models

Time of exposure should be incorporated into the analysis of data, where possible. However, different time-dependent regression models are available, each with its own assumptions, so criteria are needed for the appropriate selection from the available mechanistic and empirical models. Best reasonable fit should be the most important criterion for model selection. Also the model, whilst being as complex as necessary, should be as simple as possible. Time-dependent models are not necessarily more complex than static models (there can be fewer parameters). It may be that empirical models are less biologically consistent than mechanistic models (e.g. increases in survival may occur over time). Mechanistic models should be favoured if they fit the data; empirical models should be used as a fallback.

In the short to medium term there is no need to modify test protocols in major ways. However, minor modifications may produce a better result. It may also be useful to collect data at intermediate time intervals in some tests, such as the fish growth test, although there are more cost implications in this proposal.

6. Concerns about moving away from the NOEC

Concerns were expressed about deciding to abandon the NOEC before alternative methods have been identified and evaluated in regard to their implications for ecotoxicity test designs. It is possible that specific methods will be needed for each species and ecotoxicological endpoint. Where several methods exist, guidance for their selection for regulatory use would be required.

WORKSHOP CONCLUSIONS AND RECOMMENDATIONS

1. The NOEC should be phased out as a summary of toxicity.

2. The OECD should move towards a regression-based estimation procedure in which, as a bare minimum, the following should be reported: model parameters plus measures of error and goodness of fit; $EC_{x,\,t}$ (EC_x at time *t*), important biological parameters; parameters describing the time course of effects.

3. Time should be incorporated in the analytical procedures for OECD toxicity test data and experimental designs should be optimised for estimation of EC_x at the last time interval (where this is both relevant and cost-effective).

4. If different models are equivalent and give adequate fits to data, and if assumptions are valid, mechanistic models are preferred over empirical models.

5. Procedures for the collection of data through time should be included in test protocols that are developed or updated in the future (where this is relevant and cost-effective).

6. The OECD should initiate a study of the time-dependent regression models that are available (both mechanistic and empirical) and which of these, if any, best meets the OECD's needs. This study should include discussion of the following:

   - which statistical summaries are robust and should be reported to regulatory authorities;

   - other biological estimates or parameters that should be reported;

   - a sensitivity analysis of test design to justify an appropriate EC_x (data should be reanalysed to determine the precision associated with low EC_x values and the optimal experimental design);

   - whether a different *x* is required for different tests because of the different levels of precision achievable in each, and because an effect of magnitude of *x* may have a different biological importance for different endpoints or tests;

   - an analysis of the advantages and disadvantages to risk assessment of a move to time-dependent regression approaches for different organisms and endpoints.

7. A steering group should be set up to direct the mathematical, statistical and biological work required to take the workshop recommendations forward. This group should include representatives from the appropriate scientific and regulatory communities. The group could gather data from the testing community via a questionnaire.

8. The report of this workshop, including the background documents and the 1993 review report by Simon Pack, should be published as an OECD General Distribution Document.

GLOSSARY

ANOVA/NOEC approach

The approach whereby the mean response at each concentration is compared with the untreated mean via some statistical test. The test may be a parameter analysis of variance followed by a multiple comparison test, or may be a non-parametric test. The highest concentration that is not significantly different from the untreated is then designated as the No Observed Effect Concentration, or NOEC.

Benchmark concentration

The benchmark concentration (BC) is defined as the statistical lower confidence limit on a concentration which produces some predetermined increase in response rate compared to the untreated control. In other words, it is the lower confidence limit on an EC estimate.

Empirical vs. mechanistic modelling

Frequently a distinction is made between empirical and mechanistic models. An empirical model is one described by a family of functions which is sufficiently flexible that a member of the family fits the data well. A mechanistic model, on the other hand, is described by a family of functions that is deduced from the mathematics of the mechanism that produced the data. In reality, we find that the distinctions between the two types of model are blurred. Models that were originally thought of as mechanistic are often based upon such oversimplified assumptions that they are little more than empirical. Some families of functions are unashamedly empirical, for example polynomial and spline functions. Some models commonly used in ecotoxicology such as the logistic and Gompertz functions, which have known nonlinear behaviour built into them and which have physically meaningful parameters, are better thought of as empirical. At the other extreme are truly mechanistic or “biological” models which have a biological basis and biologically interpretable parameters. Such models should be used with caution. With the current poor state of knowledge about the vastly complex web of biology, biochemistry, nutrition and environment in which biological change is embedded, such biological bases are at best a crude approximation.

Hormesis models

Regression models fitted to dose-response data are generally monotonic, reflecting an ever-increasing adverse effect with increasing dose. Problems arise, however, when we come across effects which seem to contradict this expected monotonicity. A particular example of this is hormesis, in which low doses of a substance appear to stimulate an apparently beneficial response in the test organism even though larger concentrations lead to a toxic effect.

Static vs. time-dependent

A time-dependent model is one in which the response variable is deemed to be a function of both concentration and time. A static model is one in which the response variable is a function of concentration alone.

Threshold models

Threshold models are based upon the supposition that there exists a dose at, and below, which a substance produces no toxic effect and above which an increasing effect occurs. Threshold models can be thought of as two models joined together at a point. One part is a horizontal line describing the level of background – or untreated – response; the other describes an increasing effect with increasing dose. The two parts join at the EC_0 or threshold dose. This is the maximum dose at which the response is equal to the background and can be included in the model specification as a parameter and so be estimated directly.

ANNEX 1: LIST OF PARTICIPANTS

AUSTRIA

Norbert Bornatowicz
Österreichisches Forschungszentrum Seibersdorf
A-244 Seibersdorf
Tel: 43-1-2254-780-3540
Fax: 43-1-2254-780-3653
E-mail:bornatowicz@zdfzs.arcs.co.at

Britta Grillitsch
Veterinärmedizinische Universität Wien,
Josef Baumanngass 1,
A-1210 Vienna
Tel: 43-1-25 077-4604/1
Fax: 43-1-25 077-4790
E-mail:britta.grillitsch@vu-wien.ac.at

BELGIUM

Katrien Delbeke
LISEC
Craenevenne 140
B-3600 Genk
Tel: 32-89-36-2791
Fax: 32-89-35-5805

Colin Janssen
Lab for Biological Research in Aquatic Pollution
University of Ghent
Plateaustraat 22
B-9000 Ghent
Tel: 32-9-264-3775
Fax: 32-9-264-4199
E-mail: colin.jansen@rug.ac.be

Isabelle Halleux
ISSEP
Rue du Chera 200
B-4000 Liège
Tel: 32-4-2527150
Fax: 32-4-2424665

CANADA

Glen Atkinson
Attn. Mark Lewis
Commercial Chemicals Evaluation Branch
Environment Canada
Place Vincent Massey, 14th floor
351 St. Joseph Blvd.
Hull, Quebec K1A OH3
Tel: 1-613-232-4621
Fax: 1-819-953-4936
E-mail: g-f.atkinson@sympatico.ca

Peter Delorme
PMRA
Tupper Building
2250 Riverside Drive
Ottawa, Ontario K1A 0K9
Tel: 1-613 736-3729
Fax: 1-613 736-3710
E-mail: pdelorme@pmra.hwc.ca

Dwayne Moore
The Cadmus Group
411 Roosevelt Avenue, Suite 204
Ottawa, Ontario K2A 3X9
Tel: 1-613-761-1464/1568
Fax: 1-613-761-7653
E-mail: moored@ibm.net

DENMARK

Claus Hansen
Danish EPA, Pesticides Division
Strandgade 29
DK-1401 Copenhagen
Tel: 45-32-66-0100
Fax: 45-32-66-0479

Helle Holst
DTU, Dept. for Mathematical Modelling
Building 321
DK-2800 Lyngby
Tel: 45-45-25-3357
Fax: 45-32-66-0479
E-mail: hh@imm.dtu.dk

Gerard Jagers op Akkerhuis
National Environmental Research Institute
P.O. Box 314
Vejlsovej 25
DK- 8600 Silkeborg
Tel: 45-89-20-15-72
Fax: 45-89-20-14-14
E-mail: gja@dmu.dk

Niels Nyholm
DTU, Institute of Environmental Science and Engineering
Building 113
DK-2800 Lyngby
Tel: 45-45-25-1471
Fax: 45-45-93-2850

FINLAND

Hannu Braunschweiler
Finnish Environment Institute
P.O. Box 140
00251 Helsinki
Tel: 358-9-4030-0538
Fax: 358-9-4030-0591
E-mail: hannu.braunschweiler@vyh.fi

FRANCE

Jean-François Férard
Centre des Sciences de l’Environnement
1, rue des Récollets
57000 Metz
Tel: 33-3-8775-8180/81
Fax: 33-3-8775-8189

Eric Vindimian
INERIS
Parc technologique Alata
BP 2
60550 Verneuil en Halatte
Tel: 33-4455-6827
Fax: 33-4455-6655
E-mail: ineris@ineris.fr

GERMANY

Arno W. Lange **(Workshop Chairman)**
Umweltbundesamt
Bismarckplatz 1
Postfach 330022
D-14191 Berlin

Tel: 49-30-8903 3110
Fax: 49-30-8903 3903
E-mail: arno.lange@uba.de

Sabine Martin
Umweltbundesamt
IV 2.4
Postfach 330022
14193 Berlin

Toni Ratte
Rheinisch-Westfälische
Technische Hochschule Aachen
Lehrstuhl für Biologie V
Worringerweg 1
52056 Aachen

Tel: 49-241-806-680
Fax: 49-241-888-8182
E-mail: ratte@rwth-aachen.de

Martin Streloke
BBA
Messeweg 11/12
D-38104 Braunschweig

Tel: 49-531-299-3609
Fax: 49-531-299-3005

ITALY

Silvia Marchini
Laboratorio di Tossicologia Comparata e Ecotossicologia
Istituto Superiore di Sanità
Viale Regina Elena, 299
00161 Rome

Tel: 39-6-4990-2786
Fax: 39-6-4440-140

NETHERLANDS

Jacques Bedaux
Department of Theoretical Biology
Vrije Universiteit
De Boelelaan 1087
1081 HV Amsterdam

Tel: 31-20-444-7128
Fax: 31-20-444-7123
E-mail: bedaux@bio.vu.nl

Rinus Bogers
Notox bv.
P.O. Box 3476
5203 DL s’Hertogenbosch

Tel: 31-73-641-9575
Fax: 31-73-641-8543

Cees J. van Leeuwen
Quality Division/Risk Assessment and Environmental
Ministerie van VROM
Postbus 30945
NL-2500 Gravenhage
Tel: 31-70-339-4943
Fax: 31-70-339-1314

NORWAY

Erlend Spikkerud
Landbrukstilsynet
Norwegian Agricultural Inspection Service
P.O. Box 3
N-1430 ÅS
Tel: 47-64-94-4400
Fax: 47-64-94-4410

SPAIN

Enrique Andreu Moliner
Departamento de Biología Animal
Laboratory of Ecotoxicology
Universidad de Valencia
Dr Moliner, 50
E-46100-Bujasot (Valencia)
Tel: 34-6-386-4676
Fax: 34-6-386-4372
E-mail: enrique.andreu@uv.es

SWEDEN

Björn Dahl
Astra-Hässle AB
S-431 83 Mölndal
Tel: 4631-776-2950
Fax: 4631-776-3787
E-mail:bjorn.dahl@hassle.sc.astra.com

Lars Lindqvist
National Chemicals Inspectorate
P.O. Box 1384
S-171 27 Solna
Tel: 468-730 6836
Fax: 468-735-7698
E-mail: larsl@kemi.se

SWITZERLAND

Roland D. Fisch
Ciba-Geigy AG
Mathematical Applications/IS 2.4
R-1008.Z230
CH-4002 Basel
Tel: 41-61-697-6452
Fax: 41-61-697-8973
E-mail: wrfi@chbs.ciba.com

UNITED KINGDOM

Mark Crane
Division of Biology
School of Biological Sciences
Royal Holloway University of London
Egham, Surrey TW20 0EX
Tel: 44-1784-443372
Fax: 44-1784-470756
E-mail: m.crane@rhbnc.ac.uk

John S. Fenlon
Horticulture Research International
Wellesbourne, Warwick CV35 9EF
Tel: 44-1789-470382
Fax: 44-1789-470552
E-mail: john.fenlon@hri.ac.uk

Andrew Riddle
Zeneca Ltd.
Brixham Environmental Laboratory
Freshwater Quarry
Brixham, Devon TQ5 8BA
Tel: 44-1803-882882
Fax: 44-1803-882974
E-mail: riddle@bx1vax.zeneca.com

Tim Sparks
Institute of Terrestrial Ecology
Monks Wood, Abbots Ripton
Huntingdon, Cambridgeshire PE17 2LS
Tel: 44-1487-773381
Fax: 44-1487-773467
E-mail: ths@wpo.nerc.ac.uk

UNITED STATES

Richard G. Clements
US EPA
Office of Pollution Prevention and Toxics (7403)
401 M Street S.W.
Washington DC 20460
Tel: 1-202-260-5270
Fax: 1-202-260-1283
E-mail:clements.dick@epamail.epa.gov

David Farrar
US EPA
Office of Pollution Prevention and Toxics (7403)
401 M Street S.W.
Washington DC 20460
Tel: 1-703-305-5721
E-mail: farrar.david@epamail.epa.gov

Michael C. Newman
University of Georgia
P.O. Drawer E
Aiken, South Carolina 29802
Tel: 1-803-725-5746
Fax: 1-803-725-3309
E-mail: newman@sre1.edu

BIAC (Business and Industry Advisory Committee to the OECD)

Michael C. Harrass
Amoco Corporation
Mail Code PO619U9
130 E. Randolph Drive
Chicago, Illinois 60601
United States
Tel: 1-312-856-5116
Fax: 1-312-856-7584
E-mail:mjharrass@amoco.com

Roland Maisch
BASF AG
DUU/OO - Z 570
D-67056 Ludwigshafen
Germany
Tel: 49-621 60 58041
Fax: 49-691 60 58043
E-mail:roland.maisch@duu.x400.basf-ag.de

Joanna Jaworska Tel: 32-2-456-2076
Environmental Toxicologist Fax: 32-2-456-2845
N.V. Procter & Gamble ETC E-mail: jaworska.j@pg.com
Temselaan 100
B-1853 Strombeek-Bever
Belgium

ECETOC (European Centre for Ecotoxicology and Toxicology)

Roger Van Egmond Tel: 44-151-471-3917
Unilever Research / Port Sunlight Laboratory Fax: 44-151-471-1847
Quarry Road East
Bebington
Wirral Merseyside L63 3JW
United Kingdom

Kathleen M. Stewart Tel: 44-1803-882882
ZENECA Fax: Fax: 44-1803-282974
Brixham Environmental Laboratory
Freshwater Quarry
Brixham, Devon TQ5 8BA
United Kingdom

Lisa Tattersfield Tel: 44-151-373-5019
Shell Research Ltd. Fax: 44-151-373-5845
Thornton Research Centre E-mail: l.tattersfield@msmail.trctho.simis.com
P.O. Box 1
Chester CH1 3S11
United Kingdom

GCPF (Global Crop Protection Federation)/
ECPA (European Crop Protection Association)

Kees Romijn Tel: 33-4 9294-3453
Rhône-Poulenc Agro Fax: 33-4 9365-3924
355 rue Dostoievski
BP 153
F-06903 Sophia Antipolis
France

STEERING COMMITTEE OF THE WORKSHOP

Pascal Isnard Tel: 33-7205-2568
Rhone-Poulenc Industrialisation Fax: 33-7205-2350
C.R.I.T.
24, avenue Jean Jaurès
69153 Decines Charpieu Cedex
France

Gerhard Joermann
BBA
Messeweg 11/12
D-38104 Braunschweig
Germany
Tel: 49-531-299-3613
Fax: 49-531-299-3005
E-mail: g.joermann@bba.de

Bas Kooijman
Vrije Universiteit
de Boelelaan 1087
1081 HV Amsterdam
Netherlands
Tel: 31-20-444-7130
Fax: 31-20-444-7123
E-mail: bas@bio.vu.nl

Reinhard Meister
TFH Berlin
Fachbereich 2
Luxemburger Str. 10
13353 Berlin
Germany
Tel: 49-30-450-2213
Fax: 49-30-450-42011
E-mail: meister@tfh-berlin.de

José Tarazona
CISA-INIA
28130 Valdeolmos
Madrid
Spain
Tel: 34-1-620-2300
Fax: 34-1-620-2247

Les Touart
US EPA
OPP (7507C)
401 M Street, SW
Washington, DC 20460
United States
Tel: 1-703-305-6134
Fax: 1-703-305-6309
E-mail: Touart.les@epamail.epa.gov

Peter Chapman
Zeneca Agrochemicals
Jeallott"s Hill Research Station
Bracknell
Berkshire RG12 6EY
United Kingdom
Tel: 44-134-441-4694
Fax: 44-134-441-4853
E-mail:Peter.p.f.chapman@gbjha.zeneca.com

Simon Pack
Procter & Gamble Pharmaceuticals
Lovett House, Lovett Road
Staines
Middlesex TW18 3AZ
United Kingdom
Tel: 44-1784-49-5391
Fax: 44-1784-49-5093
E-mail: packs@pg.com

OECD

Nicky Grandy
OECD Environmental Health and Safety Division
2, rue André Pascal
75775 Paris Cedex 16
France
Tel: (33-1) 45-24-16-76
Fax: (33-1) 45-24-16-75
E-mail: nicola.grandy@oecd.org

Marie-Chantal Huet
OECD Environmental Health and Safety Division
2, rue André Pascal
75775 Paris Cedex 16
France
Tel: (33-1) 45-24-79-03
Fax: (33-1) 45-24-16-75
E-mail: marie-chantal.huet@oecd.org

Herbert Koepp
OECD Environmental Health and Safety Division
2, rue André Pascal
75775 Paris Cedex 16
France
Tel: (33-1) 45-24-76-19
Fax: (33-1) 45-24-16-75
E-mail: herbert.koepp@oecd.org

ANNEX 2: COMPOSITION OF THE WORKING GROUPS

	Working Group A	**Working Group B**	**Working Group C**
Chair	John Fenlon (UK)	Kees Romijn (GIFAP)	Michael Newman (US)
Rapporteurs	Helle Holst (Den) Les Touart (US)	Jacques Bedaux (Neth) Dwayne Moore (Can)	Colin Janssen (Bel) Gerhard Joermann (Ger)
Austria		Britte Grillitsch	Norbert Bornatowicz
Belgium	Katrin Delbeke	Isabelle Halleux	*Colin Janssen, rapporteur*
Canada	Peter Delorme	*Dwayne Moore, rapporteur*	Glen Atkinson
Denmark	*Helle Holst, rapporteur*	Niels Nyholm	Claus Hansen Gerard Jagers
Finland			Hannu Braunschweiler
France	Eric Vindimian	Jean-François Férard	
Germany	Martin Streloke	Sabine Martin Toni Ratte	
Italy		Silvia Marchini	
Netherlands	Rinus Bogers	*Jacques Bedaux, rapporteur*	Kees van Leeuwen
Norway	Erlend Spikkerud		
Spain	Enrique Moliner		
Sweden	Lars Lindqvist	Björn Dahl	
Switzerland	Roland Fisch		
UK	Mark Crane *John Fenlon, chair*	Andrew Riddle	Tim Sparks
USA	David Farrar	Richard Clements	*Michael Newman, chair*
BIAC	Roland Maisch	Michael Harrass	Joanna Jaworska
ECETOC	Roger van Egmond	Lisa Tattersfield	Kathleen Stewart
GCPF/ECPA		*Kees Romijn, chair*	
Steering Committee	Bas Kooijman (Neth) *Les Touart (US),* r*apporteur*	Simon Pack (UK) José Tarazona (Spain)	Peter Chapman (UK) Pascal Isnard (France) *Gerhard Joermann(Ger) rapporteur* Reinhard Meister (Ger)

ANNEX 3: REPORT OF WORKING GROUP A

Chairman: John Fenlon (United Kingdom)

Rapporteurs: Helle Holst (Denmark) and Leslie Touart (United States)

Session 1: Why are tests performed and what do we want from the results?

Why do we have ecotoxicity tests?

The group focused its discussion on aquatic ecotoxicity tests within the OECD framework. The OECD is concerned with developing harmonized test guidelines generally for pesticides and industrial chemicals. These tests are used mainly to derive data for regulatory purposes (e.g. for classification and risk assessments within notification and registration schemes), in order to predict possible effects of the tested substance on the aquatic environment. It is worth noting, however, that other types of risk assessments use/require different testing procedures or guidelines (e.g. bioassays for monitoring or for site-specific assessments using contaminated sediments, soils or wastewater).

In consideration of these (mainly regulatory) purposes of OECD-type tests, the group recognized that:

- Standardization of the test design is essential for the mutual use of such data by different countries, and for the use of the same data for both classification and risk assessment.

- Some flexibility in the designs is also necessary, in order to cope with the vast variety of different substances and their physical-chemical properties.

With respect to possible changes in test designs, e.g. as a consequence of changes in the statistical analysis, the group also stated that:

- The OECD should then develop methods for the continued use of existing data as well. Repetition of tests should be avoided, where possible.

What kind of information do we need from ecotoxicity tests?

Information from laboratory testing is used to predict the possibility of effects in the real environment. This extrapolation involves a high degree of uncertainty due to various causes. However, since field testing is not possible on the necessary scale and does not provide unambiguous results, laboratory testing is necessarily being used for decision-making. Therefore, it is generally agreed that:

- We must continue to improve testing and analytical techniques to move toward reducing uncertainty.

This involves two aspects: *more use of the generated data*, and *better analysis*.

As to the *use made of the data*, the existing tests already generate a lot of information but frequently only one data point is ultimately expressed (e.g. NOEC). This may be sufficient in cases where only a relative measure of toxicity is needed (e.g. priority setting, classification, limit tests to identify substances of very low toxicity). Risk assessments (prediction of possible effects) obviously require better/absolute measures of toxicity and refined analysis. The group agreed that:

⇒ Results from tests providing more detail about effects other than NOEC approaches are more useful in risk assessments.

As to *improving the analysis of test results*, a major issue in regard to reducing uncertainty is the introduction of a measure of precision concerning the test endpoint. Unlike classification schemes, risk assessments do consider a precision component. The group concluded that:

* New analytical techniques which provide for such a measure of precision and/or which reduce the extent of the necessary extrapolation are regarded as giving better results, which is generally perceived as an improvement.

* The degree of accuracy needed for each endpoint, and the use of this information in risk assessments, need to be discussed further.

* If new analytical techniques are to be introduced into the OECD Test Guidelines, the implications of the recommended technique(s) for existing test designs and for the continued use of existing data need to be evaluated.

It should also be recognized that the perception of risk may vary, and therefore some countries may require more fixed criteria for using the endpoint in a risk assessment.

Are we happy with current practices? What should tests look like in future?

The group identified several major shortcomings of the current testing practice:

- **Waste of data/imprecise results**

* ANOVA-type determination of the NOEC (i.e. by comparing control and one treatment with hypothesis tests) does not use information from all the other treatment levels (i.e. the slope of the dose-response curve).

* There is no measure of precision of the NOEC.

* The NOEC itself is imprecise, because it can only be one of the test concentrations and because the power of the statistical tests frequently does not allow for detecting considerable effects (up to 20 per cent was mentioned). Thus, bad testing (high variability of controls) is rewarded, and the NOEC as one of the tested concentrations is subject to decisions of the study director (the chosen concentrations and their spacing). This is regarded as scientifically inappropriate.

⇒ For all these reasons, the group concluded that the test design and/or the statistical analysis should be improved.

- **Ineffective use of test animals**

∗ By not using much of the data (see above), test animals are used inefficiently. This is not acceptable and needs to be improved.

- **Extrapolation needed**

∗ Both the NOEC and the EC_x correspond to a standardized exposure time, species, and laboratory condition. Hence, an extrapolating factor is needed for their use in risk assessments. With mechanistic modelling, the need for part of this factor may be reduced (e.g. the part which accounts for the extrapolation from standardized to unlimited exposure time).

- **Lack of time component**

There was agreement that more information on the time component (effect build-up over time) and more use of such data are required. This was stated for both acute and (sub)chronic tests. Such dynamic information is an important aspect in risk assessments, e.g. for evaluating the probability and extent of effects (especially when using time-dependent fate data), for evaluating the relative risks for different effects (e.g. growth reduction versus reproduction impairment), and for risk/benefit analyses.

- **Further issues to consider:**

Perception of the NOEC: Most regulators do not use NOEC values without some further interpretation of the data. One approach is comparing the NOEC with an EC_x, e.g. the EC_{10}. An NOEC > EC_{10} should then be used only carefully and with low confidence. However, the NOEC is frequently misinterpreted as a true no-effect level, especially by the lay public and by risk managers where there are more decision-making levels. As to the US, participants stated that EPA is somewhat “happy” with the NOEC in the context of its regulatory use. In EPA evaluations, it usually represents low-level effects which are not identified as statistically significant. While acknowledging the scientific shortcomings and frequent confusion with a no-effect level, the proper interpretation of other endpoints (EC_x, NEC) by risk managers and the public would also be a point of concern.

Revised test design: In case of a move towards regression analysis, several issues would need to be addressed in detail:

∗ The optimized test design may need to be different for different organisms and endpoints of concern.

∗ The same applies to the value of *x* in EC_x, due to different natural variability of endpoints like growth, reproduction, mortality with several organisms. The value of *x* and the chosen statistical technique are also likely to have implications for the number and spacing of test concentrations.

∗ Further, any change in the analytical technique should be accompanied by discussion and recommendations on the use of existing data in the future and on the parallel use of NOEC and the new measure(s) during a transition period.

Session 2: Review and comparison of the different approaches to data analysis

The group first focused on the basic comparison between NOEC (hypothesis testing) and estimation procedures with regard to regulatory tests.

Should the NOEC still have a role in future testing or should we move away from it?

In reviewing this issue, the group collated the following views:

pro NOEC	**contra**
* easy to understand * can often be determined * high confidence intervals using regression * many regression models to choose between * It is untrue to say that only control/NOEC data are used in NOEC (ANOVA incorporates variance and degree of freedom from all data, Williams' test considers additional dose-related information).	* easy to misunderstand * is often misused or inappropriate * no confidence intervals or other measure of precision with NOEC * NOEC also can depend on the choice of model or statistical test. * NOEC is not a sound and reasonable measure, for the reasons outlined in the background papers by S. Pack and P. Chapman and in the 1993 review report by S. Pack.

In conclusion, the group reached consensus on the following recommendation:

The OECD should move from the NOEC towards a regression-based (estimation) analysis of aquatic ecotoxicological data.

The group also addressed briefly the general implications of a possible change towards regression or mechanistic modelling:

* Range-finding tests for the optimal choice of test concentrations are necessary at any rate, regardless of the analytical technique.

* Test designs would need to be optimized for static regression analysis (e.g. more test concentrations, fewer replicates). Some modifications would also be needed for the dynamic approaches (time component), although existing data can already be used for modelling in DEBtox, for example. However, some changes would provide for better results.

- Guidance would need to be developed for the selection of the best model(s) (both static and dynamic) and for special cases (poor dose selection/ill-conditioned data, etc.) where model fitting fails.

- Procedures for the use of existing data sets would be necessary.

- The proper use of the additional information in risk assessments should be discussed.

Which other analytical technique could replace the NOEC?

During an animated discussion about empirical versus mechanistic modelling, one view was that many empirical models were contrary to biological knowledge and frequently inconsistent. Another view, in the minority, was that an empirical, best-fit approach was better. It was further pointed out that empirical regression methods can also incorporate time-dependent hazard/survival data, and that an EC_0 is model dependent and can be present (or not) in both empirical and mechanistic models. While doubts were expressed as to whether a move to full-scale mechanistic modelling could be achieved in one step, the group recommended that:

> ⇒ **Dynamic (time-based) components should be incorporated into the regression models. The collection of time course data should be extended.**

Which statistical measure(s) should be reported?

All raw data would be available to regulatory agencies. However, raw data normally are not used for decision-making without analysis in some form, presently through determining an NOEC. For reasons of transparent and consistent decision-making by agencies and for planning purposes by industry, such “reference point(s)” would be needed for risk assessment schemes in future. There was, however, a clear split of the group on whether NEC/EC_0 values should serve this purpose (in a vote, five members indicated they had problems with this concept and nine had no such reservations, with two abstentions).

Without agreement on that issue, the group nevertheless concluded that:

- The form and parameters of the regression model should be reported together with confidence statements.

- A general model could probably be used for many datasets.

- If the OECD decides to use mechanistic modelling, the EC_0 should be reported.

- With the EC_x approach, different values of *x* would probably be needed for different test procedures/endpoints, both in terms of matching up with the NOEC and of the achievable sensitivity. The value(s) of *x* could either be chosen to correspond to the current level(s) of possible effects at the NOEC, or could be determined by using a sensitivity analysis of the optimized test design.

- For now, values of EC_5 by increments of 5 to EC_{25} could be determined routinely.

Further, to facilitate the decision on which particular model(s) could be used in future, the group recommends that:

- **The OECD should undertake a study with existing (ring test) data to compare different types of dynamic regression models.**

In conclusion, the group reached consensus on the following recommendations:

- **The OECD should move from the NOEC towards a regression-based (estimation) analysis of aquatic ecotoxicological data.**

- **Dynamic (time-based) components should be incorporated into these regression models. The collection of time course data should be extended.**

- **The OECD should undertake a study with existing (ring test) data to compare different types of dynamic regression models.**

ANNEX 4: REPORT OF WORKING GROUP B

Chairman: Kees Romijn (GCPF/ECPA)

Rapporteurs: Jacques Bedaux (Netherlands) and Dwayne Moore (Canada)

Session 1: Why are the tests performed and what do we want from the results?

Why do we have ecotoxicity tests?

Ecotoxicity testing can be conducted for several purposes, which can be grouped in different ways:

A) Prediction	Hazard identification Classification and priority-setting Guidelines, criteria Product development Risk assessment Research
B) Control and monitoring	Permit compliance (monitoring/standards) Research
C) Diagnosis	Incidence reports Toxicity Identification and Evaluation (TIE) Research (identification of causal agents)

or

A) Legal/regulatory	New products Re-evaluations Site assessment Criteria, standards Permit-setting and compliance
B) Non-regulatory	Waste management Product development and safety Market-driven issues Emergency procedures Commercial issues Diagnosis
C) Research	Mode of action QSARs

Some of these applications of ecotoxicity data are beyond the scope of the workshop, which is to deal merely with substances and regulatory issues.

What kind of information do we need from these tests?

The kinds of information we may wish to obtain from these tests are listed below. The types of information used in various applications are shown in Table 1.

1) Lower levels of biological organisation:

biochemical
behaviour
survival
growth
reproduction

2) Population and higher levels of organisation:

Intrinsic rate of growth
Bioenergetic endpoints
Demographic endpoints
Richness, community structure
Community function endpoints

Data can be summarised using different parameters:

NOECs, LOECs, MATCs
EC_{50}, LC_{50}, IC_{50}
EC_x, LC_x, IC_p
LD_{50}
etc.

In the applications listed in Table 1, various extrapolation procedures are available for use with the ecotoxicity data including:

Uncertainty factors
Models
Uncertainty analysis

The above are used to extrapolate from one species to another, from laboratory to field conditions, from short- to long-term effects, and from lower to higher biological levels of organisation.

Are we happy with current practices?

Participants agreed that the current summary parameters for ecotoxicity testing (e.g. NOEC) are inadequate and that other summary parameters and analytical techniques should be investigated.

Session 2: Review and comparison of the different approaches to data analysis

The topics discussed were:

- Should we use NOEC/ANOVA?
- If not, what alternatives do we have?
- What measures do we want to report?
- How do we choose appropriate models?

Almost everybody agreed that ANOVA-type methods were not appropriate for estimating effective concentrations. However, if the aim is to test toxicity values against certain limits, ANOVA or other similar test procedures can be applied.

As a result, the following recommendations were made:

1) OECD tests should use a regression-based approach for the analysis of toxicity data.

2) For limit testing and other similar application, ANOVA-like methods can be applied.

Afterwards, the Group discussed what kind of regression models should be used. Two classifications were used: (1) static versus dynamic, and (2) empirical (or descriptive) versus theory-based (or mechanistic). Here static means that time is not incorporated in the model formulation. After a long discussion, the following recommendations could be made:

3) OECD Test Guidelines should encourage the characterization of the time dependence when appropriate data are available (see Table 2).

4) The OECD should evaluate whether data should be collected at intermediate times for tests that do not currently have this requirement in the guideline.

The group then discussed which summary measures are appropriate. This discussion resulted in the following recommendations:

5) The OECD should encourage the estimation of $EC_{x,t}$ including confidence limits for several values of *x* and, where appropriate, at different time intervals.

6) The OECD should reanalyse existing ring test data to determine the precision associated with low $EC_{x,t}$ values for different biological test systems and several models. The objective would be to determine the lowest effect values that can be estimated with reasonable confidence for each test system.

7) Experimental designs for OECD test systems should be optimised for estimation of $EC_{x,t}$ for the last time interval.

8) Work should be carried out within the OECD to determine:

- how dynamic data analyses would be used to effect better decision-making;
- whether this improvement justifies the extra resources required to collect and analyze dynamic data for different tests.

Two issues were also identified in this session. The first issue is that the spacing of treatments for precise estimation of low $EC_{x,t}$ is difficult for all time intervals. Therefore, the group recommended that:

9) The OECD should not modify experimental test designs explicitly for dynamic analyses unless the benefits justify the additional costs of collecting and analysing dynamic data.

The second issue raised concerned the choice of an appropriate low $EC_{x,t}$ value for risk assessment and other applications. Choosing such a value involves both statistical and regulatory considerations, and therefore was viewed as outside the scope of this workshop. Nevertheless, various regulatory and other agencies should examine this issue carefully.

The next topic discussed was the procedure for choosing models. Group B did not agree with the initial Group C proposal for choosing models based on the following priorities:

1. Best possible fit
2. Mechanistic
3. Empirical

An alternative proposal was made (which did not receive consensus):

If models give equivalent and adequate fits and if assumptions are valid, mechanistic models are preferred over empirical models.

Concern was expressed about relying solely on best-fit models, because the chosen model could change between time intervals and between tests. A last general remark was that models should be as simple as possible and as complex as necessary.

Table 1: Commonly used summary parameters in different applications of ecotoxicity data

Application	**NOEC** **LOEC**	**LC_{50}** **EC_{50}**	**LC_x** **EC_x**	**$LC_{x,t}$** **$EC_{x,t}$**	**Population level measures**	**Community level measures**
Classification	xx	xx				
Criteria and guidelines	xx	xx	xx			
Screening level risk assessments	xx	xx	x			
Higher tier risk assessments	x	x	xx	xx	xx	xx
Permit compliance	x	xx	xx	??		
Product development and stewardship	xx	xx	xx	??		
Diagnosis (e.g. TIE)	x	x	xx	x	x	
Research	x	x	xx	xx	xx	xx

x	sometimes used
xx	frequently used

Table 2: Potentially useful summary parameters from an ecotoxicity test. It is also critical that the model equation, estimated parameters and their standard error, model goodness of fit, and other test results be reported as appropriate.

% effect/time	**t_1**	**t_2**	**t_3**	**t_4**
5	$EC_{5,t1}$ ± 95% CI	$EC_{5,t2}$ ± 95% CI	$EC_{5,t3}$ ± 95% CI	$EC_{5,t4}$ ± 95% CI
10	$EC_{10,t1}$ ± 95% CI	$EC_{10,t2}$ ± 95% CI	$EC_{10,t3}$ ± 95% CI	$EC_{10,t4}$ ± 95% CI
15	$EC_{15,t1}$ ± 95% CI	$EC_{15,t2}$ ± 95% CI	$EC_{15,t3}$ ± 95% CI	$EC_{15,t4}$ ± 95% CI
20	$EC_{20,t1}$ ± 95% CI	$EC_{20,t2}$ ± 95% CI	$EC_{20,t3}$ ± 95% CI	$EC_{20,t4}$ ± 95% CI
25	$EC_{25,t1}$ ± 95% CI	$EC_{25,t2}$ ± 95% CI	$EC_{25,t3}$ ± 95% CI	$EC_{25,t4}$ ± 95% CI
30	$EC_{30,t1}$ ± 95% CI	$EC_{30,t2}$ ± 95% CI	$EC_{30,t3}$ ± 95% CI	$EC_{30,t4}$ ± 95% CI
40	$EC_{40,t1}$ ± 95% CI	$EC_{40,t2}$ ± 95% CI	$EC_{40,t3}$ ± 95% CI	$EC_{40,t4}$ ± 95% CI
50	$EC_{50,t1}$± 95% CI	$EC_{50,t2}$ ± 95% CI	$EC_{50,t3}$ ± 95% CI	$EC_{50,t4}$ ± 95% CI

ANNEX 5: REPORT OF WORKING GROUP C

Chairman: Michael Newman (United States)

Rapporteurs: Colin Janssen (Belgium) and Gerhard Joermann (Germany)

Summary

Recommendations

The working group recommended that the OECD should move away from the ANOVA/NOEC approach, and that future test methods should have the following qualities or improvements:

1. reduction of the possibility of biased results;

2. more focus on biological significance of endpoints;

3. more linkage to predictive, ecological or biological models allowing for the inclusion of relevant covariates where appropriate;

4. inclusion of guidance for explicit statistical techniques in OECD Test Guidelines;

5. choice of summary statistics that can be generated and interpreted by non-statisticians.

To this end, the fitting of regression models was recommended: specific theory-based models, if appropriate, should be used for each individual test.

Future approach

The working group recommended test-specific regression models, theory-based if appropriate. It was agreed that test results should include the following:

1. estimates of all model parameters, error terms, and goodness of fit;

2. a slope, if appropriate;

3. EC_x values;

4. biologically relevant parameters;

5. parameters describing the time course;

6. confidence limits for summary statistics, as is good statistical practice.

Session 1: Why are the tests performed and what do we want from the results?

Why do we have ecotoxicity tests?

The working group decided that the reasons for conducting ecotoxicity tests were:

- to measure absolute or relative (ranking of) toxicity values;
- for routine regulatory use;
- to determine the need for and the design of higher tier tests;
- to protect the environment by predicting the effects on natural populations (non-human);
- to generate point estimates that are useful for assessment.

What kind of information do we need from ecotoxicity tests?

The working group decided that the following type of information is needed:

- one or more endpoints (i.e. which is/are required for applying a classification scheme);
- time scale for effects;
- biologically significant information;
- indication and expression of reliability (accuracy, precision);
- information readily understandable by non-statisticians;
- representative of what is to be protected.

Are we happy with current statistical practices?

The working group identified the following shortcomings and problems in the currently used ANOVA/NOEC approach:

- unnecessarily high risk of false negatives;
- no statements of biological significance, only statistical significance;
- the type of data generated now will not meet future needs of risk assessment schemes;
- inferior statistical methods;
- no linkage to ecological, predictive models;

- inadequate extraction of information (e.g. only NOEC, no slope);
- NOECs lead to misunderstanding and misinterpretation;
- detailed guidance on statistical methods is missing in guidelines;
- no incorporation of covariates.

The working group recommended the following with regard to the design and performance of future tests and statistical analyses:

- Reduce the danger of false negatives and false positives.
- Develop tests that focus on biological significance.
- Make more efficient use of test information.
- Choose endpoints that are directly applicable to predictive ecological models and incorporate important covariates (e.g. time) where appropriate.
- The summary statistics/parameters chosen should be interpretable by non-experts.
- Make statistical tests an inherent part of test method development.
- Provide explicit recommendations for statistical data analysis in test methods and clear guidance on how to perform it/them. It must be possible for non-expert statisticians to perform the analyses.
- New statistical approaches should be compatible with GLP practice.
- Control variability should be taken into account in any analyses.
- A way to link new summary parameters to old endpoints is desirable.

Session 2: Review and Comparison of the Different Approaches to Data Analysis

Should the NOEC/ANOVA approach be retained?

A large majority of working group participants did not want to retain the NOEC/ANOVA approach. However, one participant expressed some concern about replacing the NOEC for the following reasons:

- uncertainty about what it will be replaced with;
- no infrastructure for moving away from NOEC (i.e. computer programs);
- NOEC has protected the environment in the past;
- possible need to develop new safety factors (i.e. risk assessment practices);
- the current stringent requirements for design, doses, random variation lead to valid NOECs.

What could the NOEC/ANOVA approach be replaced with?

The following options were considered:

1. summary parameters for dose-response models which allow assessment (e.g. regression, mechanistic, mixed regression/ANOVA approach, others);
2. improve NOEC methods (e.g. test design, statistical test method, reporting);
3. parallel reporting (both NOEC and EC_x), including qualification of existing NOECs;
4. EPA interpolation method.

The working group agreed to concentrate further discussions predominately on option 1.

What type of information (endpoints; summary statistics) do we want?

- **Empirical vs. mechanistic**

There was no clear consensus on whether models should be empirical or mechanistic. Some participants did not feel comfortable with the mechanistic models (e.g. DEBtox), as they did not completely understand them. The pros and cons depended on the various factors (e.g. theory used, best fit).

The working group did, however, propose the following logical sequence:

best (reasonable) fit → mechanistic if possible → if not, then use empirical models

Most participants were in favour of theory-based models, fit by non-linear regression.

- **Include covariates in the dose-response assessment?**

The working group agreed that models giving the best (reasonable) fit should be chosen first. Time should be included in this model, if appropriate. The inclusion of time *and* other covariates was selected as a second choice.

With respect to the inclusion of time as a covariate, the group acknowledged that, for some existing tests, time could be included without changes to the method (e.g. *Daphnia* reproduction study). However, other tests would need to be redesigned and this caused some concern.

What type of summary statistics do we want from our selected approaches?

Finally, the working group identified the type of summary statistics that should be reported:

- estimates of all model parameters completed by terms of error, goodness of fit;
- slope (with confidence limits);
- EC_x (with confidence limits);
- biological parameters relevant for models;
- parameters describing time course, if applicable.

Other summary statistics considered were:

- no effect concentration – NEC (or approximate of NEC);
- benchmark concentrations = lower confidence in EC_x.

Reservations were indicated with respect to the NEC. The fact that it is likely that a true NEC may exist for some chemicals, and not for others, gave concern as to its use as a parameter in chemical regulation.

ANNEX 6

A Discussion of the NOEC/ANOVA Approach to Data Analysis

Dr Simon Pack

Procter & Gamble, United Kingdom

Basic principle

The NOEC/ANOVA approach is to compare each test concentration against the control.

In general, the NOEC is the highest concentration that is not statistically significantly different from control. The LOEC is the lowest concentration that is significantly different from control.

There may be some ambiguity in these definitions if, for a concentration significantly different from control, there is a higher concentration which is not significantly different. Some might prefer an alternative definition that the LOEC is the lowest concentration significantly different from control, with the NOEC being the next lowest concentration.

Methodology

Start with parametric ANOVA, as described in basic statistics textbooks and as implemented in all general statistical computer packages. Data should be transformed to satisfy the assumptions (the assumptions should always be checked to validate any analysis). The ANOVA assumptions are that the residuals are, independently and identically, normally distributed with zero mean and constant variance. With parametric ANOVA of aquatic toxicology data, insufficient attention is often paid to the mean-variance relationship of the data, with the consequence that inappropriate estimates of variability are obtained, leading to incorrect inferences. Generalised linear models do not seem to be used.

Non-parametric versions of ANOVA exist that will generally be more robust without sacrificing much sensitivity. In fact, non-parametric methods could be used by default to avoid issues around violations of the assumptions.

To compare test concentrations with the control, there are many approaches (multiple comparison procedures). These take account of the multiple statistical tests by adjusting the statistical threshold for declaring significance. Dunnett’s method is perhaps the most popular in the ecotoxicology area. Williams’ method uses the assumed trend in the underlying concentration-response.

Despite the relative conceptual simplicity, there are, in fact, a myriad of variants when the different combinations of analysis and multiple comparisons are taken into account.

Experimental design

It is vital to identify what constitutes a replicate, e.g. individual fish in a tank or the tank itself. Investigation of the components of variability may show whether grouping of individuals can be reasonably made.

The number of replicates will significantly impact the sensitivity of statistical analyses, and therefore the NOEC. Increased control replication relative to each concentration will maximise sensitivity for a fixed overall number of replicates and is recommended. The control replication should be increased by a factor of √ (number of concentrations) relative to the test concentrations.

Optimising the design of ANOVA experiments from a statistical viewpoint seems to be seldom done.

Advantages of NOEC/ANOVA

Conceptual simplicity

With ANOVA, we simply check if each concentration is different from the control. The modelling assumptions are relatively weak, *i.e.* no concentration-response model is used. Distributional assumptions can largely be avoided by using non-parametric methods of analysis.

In comparison, regression methods aim to fit an empirical curve through the data points relating response to concentration. This curve need not be given any particular biological interpretation, although one is often assumed. The main objective is to model the data. No-effect concentrations (NECs), hormesis and time can be readily incorporated. The adequacy of the model is assessed from the fit of the model to the data points.

Unlike ANOVA and regression methods, the mathematical modelling approach makes explicit assumptions about the underlying processes. These assumptions then generate a concentration-response model for the data. Unknown parameters are then estimated from the data. For some models the experimenter may need to supply values for fixed constants; however, these should still really be considered as parameters. The assumptions can only be verified by looking at the adequacy of the fit of the model to the data points.

Computational simplicity

Hand calculation is easy. Formulae are given in most basic statistics textbooks. Many specialist and non-specialist (e.g. spreadsheet) programs are available for ANOVA calculations.

In comparison, regression and mathematical modelling approaches require specialist software. Some simple (non-parametric) methods have been developed for the estimation of e.g. EC_{50}s. However, these may not be sufficiently flexible for general recommendation. The more complex the assumptions and the model, the more complex the computational side will be and the more problems will be encountered.

Experimental design is straightforward.

The number of replicates needed to give any degree of sensitivity can be readily calculated if the standard ANOVA assumptions are assumed.

Optimal experimental design for the regression and mathematical modelling approaches is possible but not straightforward, and would most likely require further research.

Disadvantages of NOEC/ANOVA

The NOEC must be one of the test concentrations.

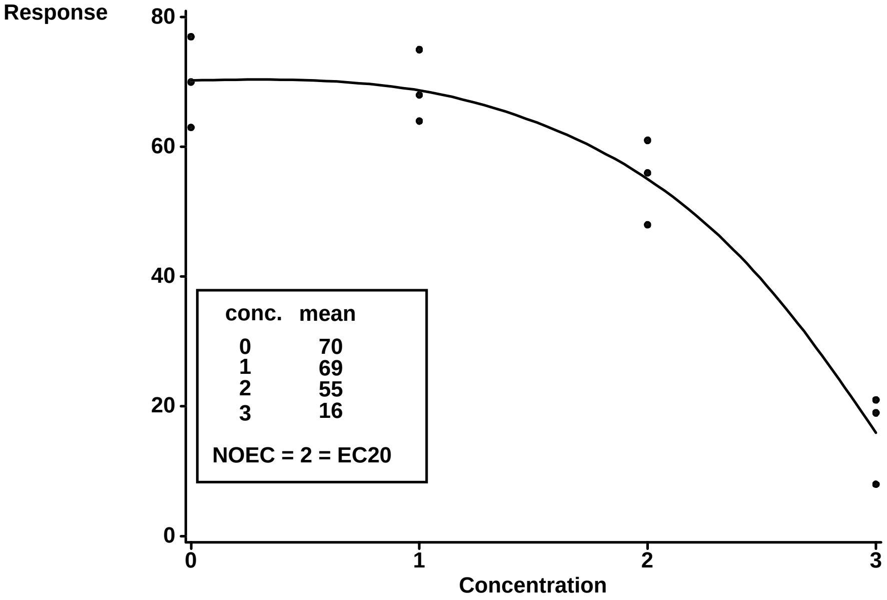
The NOEC is determined to a large extent by experimenters' choice of concentrations. Usually only a small number of concentrations are tested. Therefore the "precision" is likely to be very limited.

The NOEC is not a safe concentration.

If experimental variability is relatively high, then the sensitivity of the analyses to detect differences from control will be relatively low. This implies that only larger differences from the control can be detected. This in turn implies the NOEC may be a concentration that actually corresponds to quite large effects.

Literature supports this in practice. For example, the results of the final *Daphnia magna* reproduction ring test showed sensitivity was such that effects up to 20% could have been declared as not significantly different from control. Effect sizes at the NOECs averaged around 10% but some corresponded to effects of 20-30%.

The example below (modified from a real experiment) illustrates the problem. There is clearly an effect at the NOEC which corresponds to an EC_{20}.



NOEC obtained using Dunnett's method.
Fitted curve is 3-parameter logistic.

It is impossible to derive an estimate of the precision for the NOEC.

A power calculation can quantify the sensitivity of the analyses in terms of the difference from control that could be reasonably detected. However, this doesn't address the real problem of how accurate the NOEC is. By definition the NOEC is just one of the concentrations. If decisions are to be made on environmental safety, then it is surely important to quantify the degree of confidence.

In contrast, precision estimates are readily available from both the regression and mathematical modelling approaches.

There is no information on the concentration-response curve.

The rate of change of response with concentration may be useful in assessing how sensitive a species is. This information is not available from ANOVA/NOEC methods. Therefore valuable information is being wasted. Prediction of the effects at concentrations other than those studied is not possible.

Prediction of effects or effect concentrations is particularly simple with regression modelling and, arguably to a lesser extent, mathematical modelling approaches, since the explicit aim is to fit a model to the data points. If time is also incorporated, then predictions of effects at a given time can also be readily made.

Robustness

Parametric ANOVA is robust to moderate violations of assumptions (e.g. lack of normality). However, variations in the methodology may produce different NOEC values for the same data. The extent to which these NOECs might differ will largely depend on the spacing of the concentrations relative to the observed concentration-response. Not much seems to have been published on this.

For regression and mathematical modelling, the more complex the model used, the harder it is to validate against the data and robustness then becomes an issue.

The EC_{50} is a derived quantity and is known to be robustly estimated *i.e.* reasonably model-independent. More extreme percentiles, e.g. EC_5s, will be highly model-dependent. However, estimates of precision will be correspondingly low, reflecting the information in the data, so that confidence intervals will generally overlap for different models. Therefore, in general, simplicity in the aims and approach is recommended.

The NOEC/LOEC may not exist.

If the lowest concentration tested produces a statistically significant difference from the control, then the NOEC will not exist. If none of the concentrations is significantly different from control, then the LOEC will not exist. To obtain a NOEC/LOEC, the experiment would need to be re-run with different concentrations. The first experiment may therefore be considered to have been wasted. However, the experiment could still yield useful information on the concentration-response curve.

In contrast, both regression and mathematical modelling approaches may be able to derive useful quantities, for example EC_{50}s, from such "failed" experiments. The basic requirement would only be that the respective models can be fitted to the data.

Good experimental practice is not rewarded.

Generally, the poorer the experimental conduct the higher the variability in the data. This in turn means lower sensitivity to detect differences from control and consequently higher NOECs may result, falsely implying “greater safety”. This is completely unacceptable.

Correctly applied regression and mathematical modelling approaches will, in general, produce estimates of the parameters of interest with precision that reflects the information and variability in the data.

Other issues

NOEC/LOECs cannot be compared to EC_x values.

There have been attempts to correlate NOEC/LOECs with percentiles of the concentration-response curve, e.g. EC_{20}s. NOECs are fundamentally different from EC_xs and, as the NOEC is largely dependent on the experimental design, any correlations that have been found are almost entirely coincidental.

NOECs are not NECs.

The NOEC is not an estimate of the no-effect concentration (NEC). As already explained, the NOEC can correspond to non-zero effects and does not estimate a “safe” concentration. NOECs cannot be correlated with NECs or EC_xs.

Hybrid methods

Some authors have proposed methods aimed at overcoming the weaknesses of the NOEC. For example, Hoekstra and van Ewijk propose a two-stage procedure. Firstly, the highest concentration is found for which the effect is not estimated to be larger than 25% (the bounded-effect concentration). Then interpolation is used, from the endpoint of the confidence interval for the effect size at this concentration, to estimate the concentration giving a 1% effect (or other small value). They argue that this calculation yields a conservative estimate of a concentration giving “negligible” effects. While this may be the case, the concentration derived still suffers from the fact that an associated estimate of precision is not provided, although one could presumably be derived. Existence of the bounded-effect concentration is also not guaranteed. However, their approach does go some way to limit the dependency of the “safe” concentration on the experimental design.

References Relevant to the Debate Between the Relative Merits of NOECs and Effective Concentrations

Campbell, P.J. and S.P. Hoy. 1996. ED points and NOELs: how they are used by UK pesticide regulators. Ecotoxicology 5 (in press).

Chapman, P.F. and P.M. Chapman. A second warning: NOECs are inappropriate for regulatory use. Environ. Toxicol. Chem. (in press).

Chapman, P.F., M. Crane, J.A. Wiles, F. Noppert and E.C. McIndoe (eds.). 1996. Asking the Right Questions: Ecotoxicology and Statistics. Report of a Workshop Held at Royal Holloway University of London, Surrey, UK. SETAC-Europe.

Chapman, P.F., M. Crane, J.A. Wiles, F. Noppert and E.C. McIndoe. 1996. Improving the quality of statistics in regulatory ecotoxicity tests. Ecotoxicology 5: 1-18.

Chapman, P.M., R.S. Caldwell and P.F. Chapman. 1996. A warning: NOECs are inappropriate for regulatory use. Environ. Toxicol. Chem. 15:77-79.

Cousens, R. and C.J. Marshall. 1987. Dangers in testing statistical hypotheses. Ann. Appl. Biol. 111:469-476.

Dhaliwal, B.S., R.J. Dolan and R.W. Smith. 1995. A proposed method of improving whole effluent toxicity data interpretation in regulatory compliance. Water Environ. Res. 67: 953-963.

Dhaliwal, B.S., R.J. Dolan, C.W. Batts, J.M. Kelly and R.W. Smith. 1996. Warning: Replacing NOECs with point estimates may not solve regulatory contradictions. A fundamental change in toxicity data interpretation is warranted. Envion. Toxicol. Chem. (in press).

Hoekstra, J.A. and P.H. Van Ewijk. 1993. Alternatives for the no-observed-effect level. Environ. Toxicol. Chem. 12:187-194.

Kooijman, S.A.L.M. 1996. An alternative for NOEC exists, but the standard model has to be abandoned first. Oikos 75: 310-316.

Lacey, R.F. and M.J. Mallett. 1991. Further statistical analysis of the EEC ring test of a method for determining the effects of chemicals on the growth-rate of fish. Unpublished report of an OECD meeting of experts on aquatic toxicology, WRc, Medmenham, UK, 10-12 December, 1991.

Laskowski, R. 1995. Some good reasons to ban the use of the NOEC, LOEC and related concepts in ecotoxicology. Oikos 73: 140-144.

Leisenring, W. and L.M. Ryan. 1992. Statistical properties of the NOAEL. Regul. Toxicol. Pharmacol. 8:161-171.

Masters, J.A., M.A. Lewis, D.H. Davidson and R.D. Bruce. 1991. Validation of a four-day ceriodaphnia toxicity test and statistical considerations in data analysis. Environ. Toxicol. Chem. 10:47-55.

Moore, D.R.J. and P.Y. Caux. 1997. Estimating low toxic effects. Environ. Toxicol. Chem. 16 (4) (in press).

Noppert, F., N. Van der Hoeven and A. Leopold. 1994. How to measure no effect: towards a new measure of chronic toxicity in ecotoxicology. Workshop Report of the Netherlands Working Group on Ecotoxicology. The Hague.

Pack, S. 1993. A review of statistical data analysis and experimental design in OECD aquatic toxicology test guidelines. Shell Research Limited, Sittingbourne Research Centre, Sittingbourne, Kent, ME9 8AG, UK.

Petersen, R.G. 1977. Use and misuse of multiple comparison procedures. Agronomy Journal 69:205-208.

Skalski, J.R. 1981. Statistical inconsistencies in the use of no-observed-effect levels in toxicity testing. In: D.R. Branson and K.L. Dickson (eds.), Aquatic Toxicology and Hazard Evaluation, Fourth Conference, ASTM STP 737, American Society for Testing and Materials.

Stephan, C.E. and J.W. Rogers. 1985. Advantages of using regression analysis to calculate results of chronic toxicity tests. In: R.C. Bahner and D.J. Hansen (eds), Aquatic Toxicology and Hazard Evaluation, Eighth Symposium, ASTM STP 891, American Society for Testing and Materials.

Suter, G.W., II. 1996. Abuse of hypothesis testing statistics in ecological risk assessment. Human Ecol. Risk Asssess. 2: 331-347.

References on Threshold and Hormesis Models

Brain, P. and R. Cousens. 1989. An equation to describe dose-responses where there is stimulation of growth at low doses. Weed Research 29:93-96.

Cox, C. 1987. Threshold dose-response models in toxicology. Biometrics 43:511-523.

Ewijk, P.H. van and Hoekstra, J.A. 1993. Calculation of the EC50 and its confidence interval when subtoxic stimulus is present. Ecotoxicology and Environmental Safety 25:25-32.

Yanagimoto, T. and E. Yamamoto. 1979. Estimation of safe doses: critical review of the hockey stick regression method. Environmental Health Perspectives 32:193-199.

References on the Benchmark Concentration

Crump, K.S. 1984. A new method for determining allowable daily intakes. Fundam. Appl. Toxicol. 4:854-871.

Crump, K., B. Allen and E. Faustman. 1995. The use of the benchmark dose approach in health risk assessment. Report of the US EPA Risk Assessment Forum, EPA/630/R-94/007, February 1995.

References on Time to Response Modelling

Dixon, P.M. and M.C. Newman. 1991. Analysing toxicity data using statistical models for time-to-death: An introduction. In: Michael C. Newman and Alan W. McIntosh (eds.), Metal Ecotoxicology, Concepts and Applications. Lewis Publishers, Inc., Chelsea, Michigan.

Kooijman, S.A.L.M. 1993. Dynamic Energy Budgets in Biological Systems: Theory and Applications in Ecotoxicology. Cambridge University Press, Cambridge.

ANNEX 7

Alternatives to the NOEC Based on Regression Analysis

Peter Chapman

Zeneca Agrochemicals, United Kingdom

Effective Concentration (EC) estimation

An “effective concentration” (EC) is defined as the concentration that produces a specified size of effect relative to an untreated control. An equivalent definition applies to an effective dose, but concentration is used throughout this document.

Methodology

A typical experiment comprises an untreated control plus a number of concentrations replicated a number of times. The measurement made on each experimental unit is some form of sub-lethal response such as the weight of an organism or the number of offspring produced. A regression model is fitted to the data and, through a process known as inverse estimation, a concentration corresponding to a specified percent effect relative to the control is estimated. For example, a concentration corresponding to a 50% reduction in effect relative to the control could be estimated. Figure 1 illustrates how an EC is estimated for a typical sigmoidal dose-response curve. Confidence intervals for the EC can, and should always be, estimated.

The curve fitted will usually be empirical in nature and will not have any particular biological justification. However, experience demonstrates that many curves, such as that illustrated in Figure 1, fit ecotoxicity data well and are adequate for the purpose of estimating effective concentrations.

Experimental design

The design will usually take the form of a fully randomised or randomised complete block, although more complicated designs are possible. If it has been decided to replicate the control and the doses, it is important that replication is true replication and not pseudo-replication. (Pseudo-replication occurs when sub-samples from an experimental unit, such as measurements on individual organisms within a single housing unit, are used in the analysis of an experiment as if they are true replicates.) The number of replicates need not be the same for each dose.

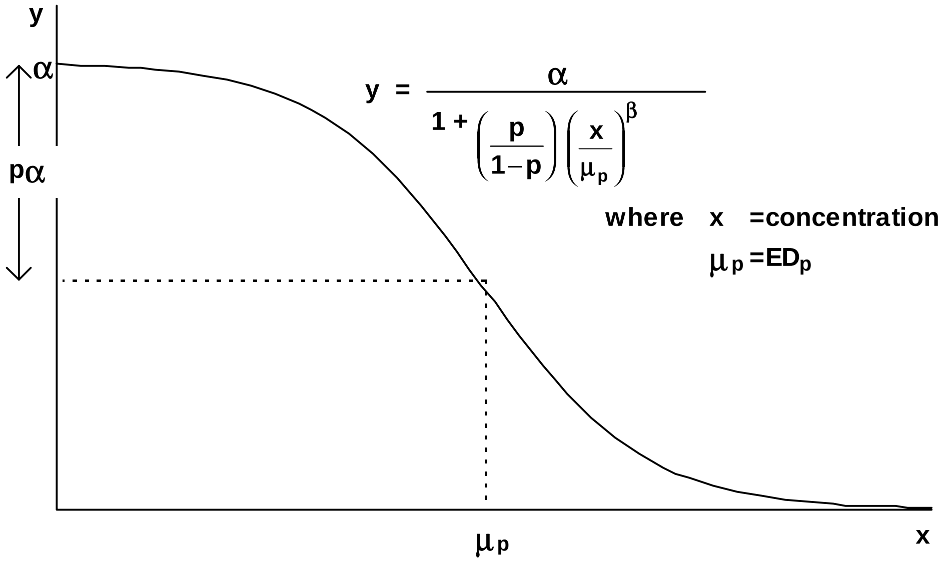


Figure 1: Concentration-response curve for the logistic regression equation

Potentially useful extensions to regression procedures

Benchmark concentration

The benchmark concentration (BC) is defined as the statistical lower confidence limit on a concentration which produces some predetermined increase in response rate compared to the untreated control. In other words, it is the lower confidence limit on an EC estimate. The BC is a relatively conservative quantity on which to base the estimate of a safe concentration, but this characteristic has found it many advocates in fields such as mammalian toxicology where the potential risk to humans must be kept very small indeed. It may however be considered too conservative for use in ecotoxicology.

The EC_0

Commonly used dose-response models, such as the logistic curve, predict non-zero effects even at very small doses, so an EC estimate cannot be regarded as an estimate of a true No Effect Concentration (NEC). It is possible, however, to modify conventional dose-response models in order to estimate an EC_0. Two such types of model which deserve serious consideration are threshold models and hormesis models.

Threshold models

Threshold models are based upon the supposition that there exists a dose at, and below which, a substance produces no toxic effect and above which an increasing effect occurs. Threshold models can be thought of as two models joined together at a point. One part is a horizontal line describing the level of background – or untreated – response; the other describes an increasing effect with increasing dose. The two parts join at the EC_0 or threshold dose (see Figure 2). This is the maximum dose at which the response is equal to the background and can be included in the model specification as a parameter and so be estimated directly.

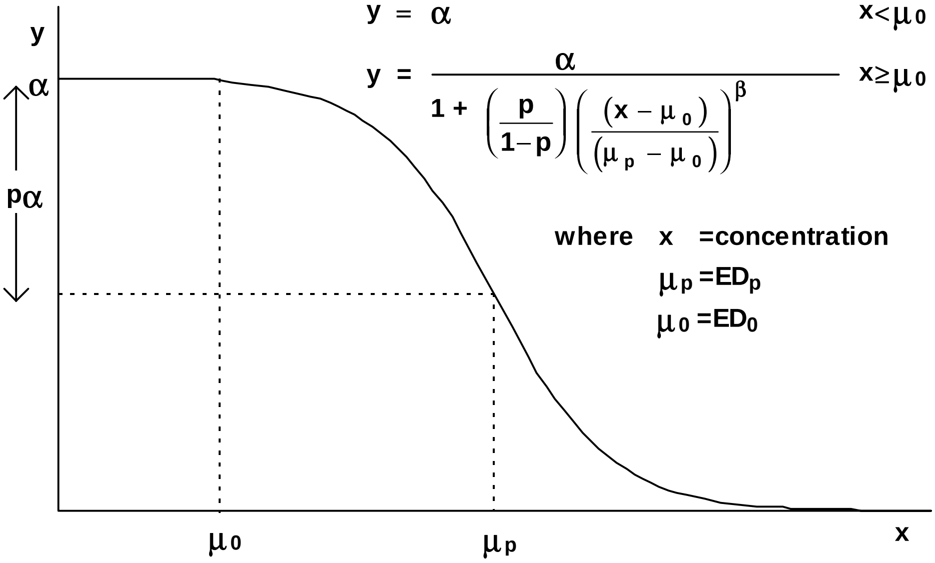


Figure 2: Calculation of a threshold concentration or EC_0.

Hormesis models

Regression models fitted to dose-response data are generally monotonic, reflecting an ever-increasing adverse effect with increasing dose. Problems arise, however, when we come across effects which seem to contradict this expected monotonicity. A particular example of this is hormesis, in which low doses of a substance appear to stimulate an apparently beneficial response in the test organism even though larger concentrations lead to a toxic effect (see Figure 3). From an ecotoxicological perspective, the point at which the response is equal to the untreated response – the EC_0 in Figure 3 – could be interpreted as the concentration at which the stimulatory effect ceases and the toxic effect begins and thus could be regarded as an estimate of the true NEC.

Time to response models

If the responses in a single ecotoxicity test are measured on a number of different days, then a sigmoidal concentration response curve can be modified to include time. One of the benefits is that EC estimates can be given for different times. An additional benefit is that time to response can be estimated as a function of concentration.

Advantages and disadvantages of EC estimates

Below is a list of advantages and disadvantages of EC estimation as an alternative to the NOEC. The list of advantages mainly draws attention to ways in which EC estimates overcome the scientific deficiencies in the NOEC. In contrast, the list of disadvantages describes some of the difficulties to be encountered in trying to estimate EC points. All of these difficulties can be overcome, although to do so may require very large experiments or a significant amount of resource in analysing data from an experiment.

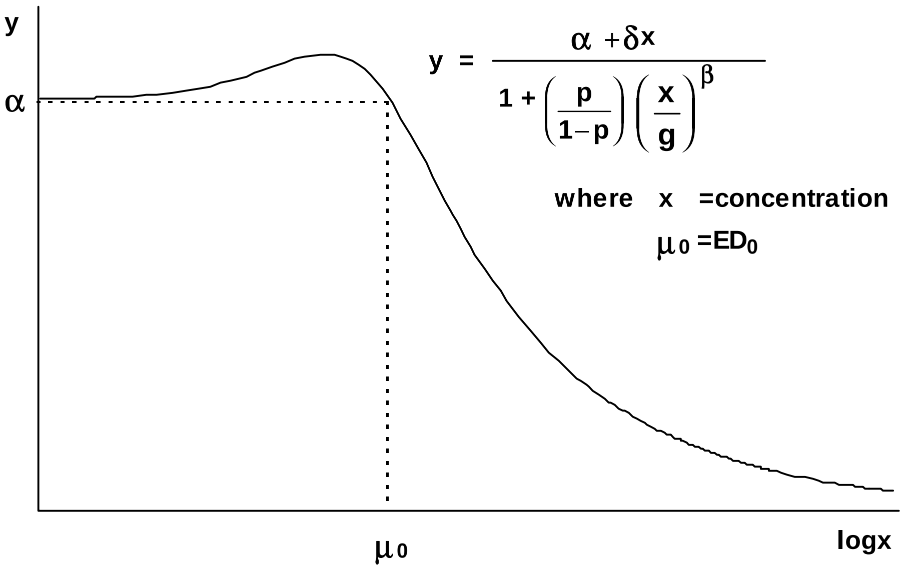


Figure 3: EC_0 estimation in the presence of stimulation of response at low concentrations (hormesis)

Advantages

1. Regression permits the estimation of effects at untested doses. In contrast, a NOEC can only be one of the doses actually used in the experiment.

2. EC estimation rewards well conducted experiments - i.e. those which are unbiased and of low variability. The greater the precision in an experiment the smaller will be the width of the confidence interval around the estimate. This implies that the lower the variability in an experiment the more likely it is that an EC estimate will be close to its true value. In contrast, highly variable results are more likely to produce EC estimates that are much larger, or much lower, than their true values.

   Standard ANOVA/NOEC evaluation, on the other hand, rewards poor experiments (i.e. high variability) with high NOEC values.

3. Because confidence intervals can always be calculated, the precision of an EC estimate can always be determined, whereas the precision on a NOEC can never be determined.

4. An EC can always be estimated, even if it is outside the range of experimental doses, although estimates far outside the range of doses should be treated with caution.

   By contrast, the NOEC is not always obtainable, either because the lowest dose gives a statistically significant effect or none of the doses do. This situation has resulted in tests having to be repeated.

5. The biological effect produced by a dose equal to the EC is not zero but it is known because it is pre-selected. This situation is to be preferred to that of the NOEC, which also produces a non-zero effect which can be quite large, but is of unknown magnitude.

6. Unlike the NOEC, the estimated EC does not depend upon the type I error rate in a significance test, nor on the choice of multiple comparison procedure.

Disadvantages

1. The choice of regression model: If one is interested in estimating an EC for a small effect size, such as an EC_{10} or EC_5, the estimate will usually depend upon the choice of model. For any single set of test results, therefore, the data analyst may have to fit a large number of models, and even then it may be difficult to decide which is best.

2. Even if one is confident that the right model has been fitted, for small effect sizes the confidence interval around the EC estimate will be relatively large.

3. Precision of EC estimates depends upon the number of test concentrations and their values. Therefore, if it is important to be able to estimate small effect concentrations, some research will need to be carried out to determine optimum selection of test concentrations. This may need to be done separately for each model.

4. Choice of effect size: If EC estimation is to replace the NOEC, then for each type of endpoint in each type of test the size of effect of interest needs to be decided upon.

5. For Benchmark Concentrations not only do the model and the effect size need to be chosen but also the confidence level used in estimating the confidence interval.

6. Both threshold and hormesis models require at least one extra parameter compared with a normal sigmoid model. This makes it more difficult to fit the model.

7. Experience suggests that confidence intervals around NEC estimates from threshold and hormesis models tend to be very wide. Thus before these models can be used some research into the optimal selection and placing of doses is required.

NOTE: For references relevant to the debate between the relative merits of NOECs and effective concentrations, see the preceding paper by Simon Pack, "A Discussion of the NOEC/ANOVA Approach to Data Analysis".

ANNEX 8

Dynamic measures for ecotoxicity[*]

S.A.L.M. Kooijman, J.J.M. Bedaux

Vrije Universiteit, Dept. Theoretical Biology, de Boelelaan 1087, NL-1081 HV Amsterdam

A.A.M. Gerritsen, H. Oldersma & A.O. Hanstveit
TNO-Toxicology, Dept. Envir. Toxicology, P.O. Box 6011, NL-2600 JA Delft

12 September 1996

Abstract

There are three required components of dynamic models for toxic effects: toxico kinetics, effects on a target parameter coupled to the internal concentration and the physiological component. The Dynamic Energy Budget (DEB) model, which is used to model the latter component, relates a change in a target parameter of a particular physiological process, such as the specific costs for growth, to an output variable, such as the cumulative number of offspring. We compare the logit/probit and the DEB-based models conceptually and numerically and conclude that the DEB-based model is more effective as an effect model. The DEB-based model solves the problem of estimating the No-Effect Concentration and provides the required information to evaluate the consequences of effects on individuals for population dynamics.

1 Introduction

In environmental risk assessments Predicted No-Effect Concentrations (PNECs), that are derived from No-Observed Effect Concentrations (NOECs) in standard single species toxicity tests, are compared with Predicted Environmental Concentrations (PECs), derived from production volumes, use patterns, and transport in the environment. The purpose of the toxicity tests, together with their designs and experimental protocols, evolved gradually towards sophistication. The analysis of the results of these tests, however, did not catch up with these changes for a long time. This document introduces a process-based approach for the analysis of toxicity tests, which has a firm rooting in biology.

The purpose of this document is to identify the aim of toxicity tests and to present the DEB-based effect model as a method to achieve this aim. The static and the dynamic methods are compared, conceptually and numerically. We tried to refrain from technical details in this paper and refer to Kooijman and Bedaux (1996) for a full discussion of the DEB-based model for toxic effects. A short introduction to the DEB-theory is given in a separate document.

[*] Background Document for the OECD Workshop on Analysis of Data from Aquatic Ecotoxicity Tests, Braunschweig, Oct, 1996

1.1 Aims of toxicity tests

The primary purpose of standard toxicity tests is to provide information about

- the maximum concentration that gives no effect on a response variable (such as survival, body growth, reproduction, population growth). This information is used to derive a level in the environment that can be considered as 'safe'.

- what effects are to be expected in the environment if these levels are exceeded (a little bit)?

- the requirement for further in-depth studies with respect to the ecotoxicity of the chemical. The priority is high if expected levels in the environment are likely to approach or exceed the level that is considered 'safe' and, in that case, the effect is likely to be substantial. The actual decision to further research depends on financial possibilities for research and socio-economic factors in industrial activity.

The second and third application of toxicity tests imply a quantification of effects as a function of the concentration. They also imply an extrapolation of the effects as observed during the tests to effects at long-term exposure and a translation to effects in the environment.

1.2 Classification of approaches

The analyses of toxicity tests can be classified into three approaches, the ANOVA method, static methods and dynamic methods.

- The ANOVA method aims to identify the highest tested concentration with a response that does not deviate from the blank on the basis of a statistical procedure: the NOEC.

- The static method quantifies effects at a standardized exposure time on the basis of concentration response models, such as the logit, probit or Weibull model, which are all very similar. The toxic effect is quantified by EC50 (LC50) or an EC*x* value for some small value of *x*. The method is called static because information about the rate at which effects build up during exposure is not used. Fixed extrapolation factors are used to predict effects at long exposure.

- The dynamic method quantifies effects as functions of the concentration *and* the exposure time. The Dynamic Energy Budget (DEB)-based model is an example of this approach. The toxic effect is quantified by two parameters: the no-effect concentration (NEC), and the killing rate (for survival) or the tolerance concentration (for other endpoints) as a measure of the effect if the NEC is exceeded.

2 The dynamic approach

The dynamic approach for the analysis of standard toxicity tests characterizes toxic effects with a No-Effect Concentration (NEC), a tolerance concentration (or a killing rate in the case of effects on survival), and an elimination rate. These parameters directly relate to long-term exposure, so no extrapolation factor is required here. They relate to changes in processes. From these three parameters, it is possible to calculate the static parameters, including the EC0 at the end of the test (therefore comparable to the NOEC), as well as the full EC*x*-time behaviour. It is not possible to calculate the dynamic parameters from the static ones, which shows that the dynamic parameters contain more information than the static ones. The dynamic approach uses information about the rate at which effects build up during exposure. It is important to realize that static and dynamic approaches are alternative analyses for the *same* toxicity tests. The approaches only differ in the type of information that is extracted from the data. Given the choice for the analysis of the data, we can try to optimize the experimental setup of the test in terms of efficiency. This is another issue that is beyond the scope of this paper.

2.1 The components of dynamic effect models

The three components that any dynamic model for effect should have are the kinetics, the effect and the physiological component, which are discussed in the next subsections. The implication being that responses are modelled as function of the concentration *and* exposure time. We have a response *surface*, rather than a concentration response relationship. For a response such as the cumulative number of offspring, this means that we include any delay of the start of the reproduction into the analysis.

2.1.1 Kinetics

The kinetics component links internal concentrations to external concentrations. The first order kinetics (also called the one-compartment model) is the simplest choice. It assumes that uptake is proportional to the external concentration and elimination is proportional to the internal concentration. If the organism grows during exposure, dilution by growth results in a deviation from first order kinetics, that should be taken into account. The DEB-based model takes uptake and elimination rates proportional to the surface area, while the DEB component (see below) specifies the growth process.

Actual measurements of time profiles for the internal concentration sometimes show deviations from first order kinetics. More-compartment models are frequently used to improve the fit, because they have more parameters. A more detailed modelling of the various uptake rates (via water and/or food), elimination rates (via water, reproduction and/or respiration), changes in fat content, and metabolic transformations, is frequently more realistic than using multi-compartment models. Some chemicals are not taken up at all, but have their effect on the outer side of the organism. Multi-compartment models should only be used if the compartments are identified and their concentrations of toxicant measured. The latter requirement applies to all kinetics models that are more complex than the first-order one. Reconstructions of complex kinetics from observed effects, rather than from measured internal concentrations, run into problems rather easily.

Since internal concentrations are not measured in standard toxicity tests, application of more advanced models for kinetics in these toxicity tests is not feasible. The alternative to refrain from dynamic modelling and apply an arbitrary extrapolation factor to arrive at predicted long-term effects is certainly not a better alternative. We consider the application of a safety factor to compensate for possibly inappropriate use of first-order kinetics to be a more promising alternative. This problem is inherent to the use of standard toxicity tests for environmental risk assessments; an in-depth scientific study for long-term effects is the only sensible alternative. Because of the financial costs of such a study, this will only be possible for a limited number of chemical compounds.

2.1.2 Effects

The effect component links effects on a target parameter, such as the volume-specific costs for maintenance or growth, to the internal concentration. The linear effect model is the simplest choice: The change in the target parameter is proportional to the internal concentration that exceeds the internal No Effect Concentration (NEC). Translation of this concept into formulae gives the following relationships for sublethal and lethal effects on a target parameter: if int.conc (*t*) $\geq$int.NEC at exposure time *t:*

$$\mathrm{par_c}(t)=\mathrm{par_0}\left(1+\frac{\mathrm{int.conc}(t)-\mathrm{int.NEC}}{\mathrm{int.tolerance\ conc.}}\right)$$

$$\mathrm{haz_c}(t)=\mathrm{haz_0}+\text{killing rate}\ \frac{\mathrm{int.conc}(t)-\mathrm{int.NEC}}{\mathrm{BCF}}$$

where BCF stands for the BioConcentration Factor (the ratio of the internal and the external concentration after long-term exposure) and haz stands for the hazard time *i.e.* the instantaneous death rate. As long as the internal concentration is less than the NEC, we have that $\mathrm{par_c}=\mathrm{par_0}$ and $\mathrm{haz_c}=\mathrm{haz_0}$. If we divide the internal concentrations by BCF we get external concentrations. The result is for conc (t) $\geq$ NEC.

$$\mathrm{par_c}(t)=\mathrm{par_0}\left(1+\frac{\mathrm{conc}(t)-\mathrm{NEC}}{\mathrm{tolerance\ conc.}}\right)$$

$$\mathrm{haz_c}(\mathrm{t}) = \mathrm{haz_0} + \text{killing rate}\ (\mathrm{conc}(t) - \mathrm{NEC})$$

where conc*(t)* has the dimensions of an external concentration but it is proportional to the internal one. Note that the NEC refers to long-term exposure. The NOEC for a particular toxicity test depends on the maximum observed exposure time *t*. It is therefore conceptually more or less comparable with the EC0.*t*, while the NEC equals the EC0∞.

The above mentioned formulae define the killing rate and the tolerance concentration; these parameters occur as proportionality coefficients in the description of effects. The killing rate has dimension 'per environmental concentration per time' and is a measure for the toxicity of a chemical if it exceeds the NEC; a toxicity measure that is independent of the exposure time. The tolerance concentration has the dimension 'environmental concentration'. The more toxic the chemical, the lower is its value. It is essential to specify the target parameter to which the tolerance concentration relates.

Five target parameters are distinguished for effects on reproduction: two for direct and three for indirect effects. Direct effects on reproduction are defined as effects on what happens with the investment into reproduction, not on the size of investment itself. One mode of action increases the energetic costs of each young, the other affects the survival of each ovum during a short sensitive period. Indirect effects on reproduction affect the investment into reproduction, not the conversion of this investment into young. Indirect effects increase the costs of growth or maintenance, or decrease the assimilative input. These three indirect effects not only reduce the reproduction, but also delay the start of the reproduction.

Consistent with the five target parameters for effects on reproduction, there are three target parameters for effects on body growth: one direct effect (the increase of the specific costs for growth), and two indirect ones: the increase of maintenance costs and the decrease of the assimilative input.

For effects on population growth of algae, we distinguish a direct effect on cell growth (by increasing the specific costs for growth), and two effects on survival (so the target parameter is the hazard rate): one effect lasts during population growth and the other only operates during a very short period after inoculation. In the latter case, effects are supposed to occur only during the transition from the culture to the experimental conditions. The sensitivity of the cells is here supposed to relate to the cell cycle. The overall effect is a delay of population growth, rather than a reduction. The uptake/elimination kinetics is assumed to be fast relative to the concentration equals the product of the BCF and the external concentration.

The linear relationships between the internal concentration and the target parameter follow from two arguments: effective molecules operate independently, and it is a simple approximation for small changes in the target parameter. The improvement of goodness of fit for large changes in the target parameter probably does not balance extra parameters. Moreover, at higher concentrations, more physiological processes are likely to be affected simultaneously. This makes that many parameters have to be introduced to capture large effects. We consider such approvements counter productive in standard toxicity tests. The argument implies that if the goodness of fit of the model to the responses is less than excellent, a higher weight should be given to responses at low concentrations.

Note that a linear relationship between the target parameter and the internal concentration, does not imply a linear relationship between the response (*i.e.* output variable) and the external concentration. These relationships work out to be sigmoid.

2.1.3 Blank physiology

For acute lethal effects we do not need a model for the blank physiology. For effects on algal population, such a model can be very simple. For effects on body growth and reproduction, however, we need a physiological component that links output variable(s) to the target as body length, cumulative number of offspring, number of surviving individuals, *etc*. The Dynamic Energy Budget (DEB) model is the simplest choice that links all essential processes: feeding, digestion, respiration, maintenance, growth, development, reproduction and ageing. It is introduced in a separate document. Many other models for the uptake and use of food have been described in literature, which typically involve many parameters. A comparative discussion of these models is beyond the scope of this paper. Since the physiological component only relates to the ecophysiology of the test species that is involved, and not to the toxicant, its applicability needs not to be studied for each individual toxicity test. It needs to be studied only once per species. The physiological component in this respect differs fundamentally from the other two components: the kinetics and the effect component.

Some chemicals, such as endocrine disrupters, might have an effect at the molecular level that does not directly relate to the energetics of the organism, but other effects will eventually translate into effects on the energetics; it is the choice for output variable (*e.g.* growth or reproduction) that directly relates to the energetics, not the molecular mode of action of the chemical. If a chemical has neither direct nor indirect effects on energetics, such as a chemical that only affects behaviour, it will have no effects in toxicity tests for effects on growth or reproduction.

Effects on survival directly affect the hazard rate, which can be studied without detailed reference to energy budgets. It is perhaps better to call the model a hazard-based model, rather than a DEB-based one; the rich structure of the DEB model only comes into play for effects on growth and, especially, reproduction. The fact that changes in the hazard rate by toxicants beautifully link up with the effect of the ageing process in the DEB model, can be considered as a happy coincidence that has little relevance for most standard toxicity tests. It becomes more relevant if the length of the toxicity test is not short with respect to the life span of the test organism. Evaluation of the consequences of effects on survival for population dynamics does involve the complete structure of the DEB model.

Effects on the population growth rate for dividing organisms such as algae only involve certain simple aspects of the DEB model, as is explained in Kooijman (1993). This is because the surface area/volume ratio only changes within a very restricted range during the cell cycle. This does not hold for animals such as water fleas and fish.

3. Examples of application of the DEB-based model

Examples of the application of the dynamic approach for toxicity tests for effects on survival, body growth, reproduction and population growth are given in Figures 1, 2, 3 and 4. All figures are composed from output files of the software package DEBtox, as provided in Kooijman & Bedaux (1996). It fits the response surface to all available data simultaneously, so the different curves in the concentration and exposure time profiles are linked and are not fitted independently.

Since the use of profile ln likelihood functions for the identifications of confidence intervals (here for the NEC) is not standard (probably because of the substantial amount of calculations a confidence level is first selected on the horizontal axis in the left panel; the threshold value for the ln likelihood function is then read off on the vertical axis, and the graph in the right panel is (mentally) intersected at this level. The intersection points represent the boundaries of the appropriate confidence interval. More than one confidence interval might result, because the ln likelihood function for the NEC can deviate substantially from a simple parabola, which is the shape that it should have if the large-sample theory for parameter estimation would apply. This is why this way of obtaining confidence sets is much more reliable than making use of the Asymptotic Standard Deviation (ASD, given in the parameter tables). The profile ln likelihood functions for the NEC in the toxicity tests on fish body growth and algae population growth are close to the expected parabola in these examples. DEBtox can also produce the numerical values of the interval estimates. Note that, generally, the NEC cannot easily be guessed from the concentration profiles, because this threshold corresponds with the EC0.*t*, while we have that NEC = EC0.∞.

The confidence intervals for the NEC in Figure 3 are not small. The main reason is in the uncertainty about the value of the elimination rate. If other information about this rate could be supplied (for instance from direct measurements, comparison with other toxicity tests or Quantitative Structure Activity Relationships), the confidence interval for the NEC can be reduced substantially.

DEBtox can also be used to test parameter values statistically and to extract all kinds of information about the effect surface, such as EC*x* values. This is not illustrated here.

4. Dynamic versus static approaches

4.1 Toxicity comparisons

Apart form comparing different models for the same data, we can and should compare data from different toxicity tests (different species of test organisms, different test chemicals, different endpoints). These comparisons yield some arguments that play a role in the comparison of the different models to the same data, and are, therefore, presented first.

4.1.1 Solubility in fat

The solubility in fat is a physical chemical property that is very relevant to the toxicity of a chemical. The linear effect model (*i.e.* the effect component of the DEB-based model), assumes that effective molecules operate independently. Since the ultimate number of molecules in an organism is directly proportional to the octanol/water partition coefficient, P_{ow}, the tolerance concentration and the NEC should be proportional to P_{ow}^{-1} and the killing rate to P_{ow}.. The latter relationship directly follows from the argument that the inverse of the killing rate is proportional to a tolerance concentration, as is obvious from the dimensions of the killing rate.

The symmetry argument states that the uptake flux depends on P_{ow} in the same way as the elimination flux depends on P_{wo}; while P_{wo} = P_{ow}^{-1}. This argument directly results in the expectation that the uptake rate is proportional to $\sqrt{P_{ow}}$ and the elimination rate is proportional to $1/\sqrt{P_{ow}}$. The logic is easily seen when we realize that the BCF equals the ratio of the uptake and the elimination rate, while it is proportional with $P_{ow.}$

The strength of these simple relationships becomes obvious if we have the results of a toxicity test with a chemical with a P_{ow} of P_1 available and wonder about the toxicity of another test chemical with a P_{ow} of P_2. The relationships tell us to multiply the elimination rate with $\sqrt{P_1 / P_2}$, and the NEC and tolerance concentration with P_1P_2. These three statistics define the complete EC*x*.time behaviour of the second chemical. These expectations can help a lot in choosing the concentrations that are to be used in a toxicity test for the second chemical. This increase in efficiency helps to reduce the financial costs of a toxicity test.

Figure 5 presents the NEC, killing rate and elimination rate as functions of the octanol/water partition coefficients for alkyl benzenes. The theoretical predictions seem to apply, but the scatter, particularly for the elimination rate, is substantial. The simplicity of the relationships of the DEB-based model parameters with the P_{ow} also helps to detect patterns in toxicity, comparing many test chemicals.

The EC50 of the logit model depends on the P_{ow} in a much more complex way, due to the fact that it depends on the standardized exposure time. Chemicals with a sufficiently small P_{ow} reveal their toxic properties fully during the toxicity test, but chemicals with a large P_{ow} do not and the apparent toxicity will increase if the test would be extended. This can be understood on the basis of the relationship between the elimination rate and the P_{ow} as mentioned above. The same problem applies to the NOEC. The allometric model EC50=a$(P_{ow})^b$ is usually fitted, but it is only based on empirical arguments.

Survival, Hazard model			ASD	Correlation coefficients		
Blank mortality rate	2.296e-011	d^{-1}	0.000			
No-effect concentration	190.2	$ug\ l^{-1}$	0.783	0.000		
Killing rate	0.009304	$l\ ug^{-1}\ d^{-1}$	0.001	0.000	0.081	
Elimination rate	7.019	d^{-1}	2.510	-0.000	0.134	-0.700
Deviance	31.41					

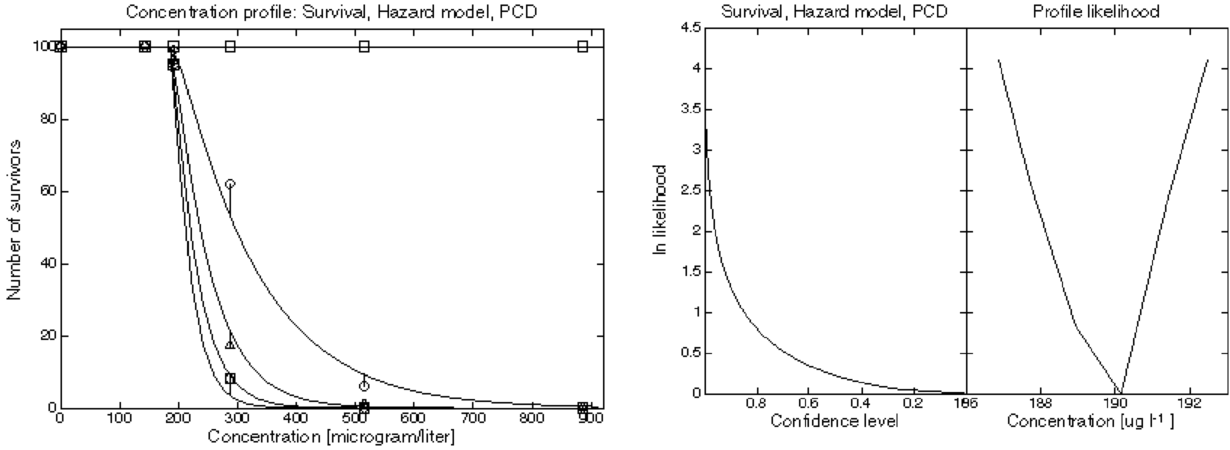
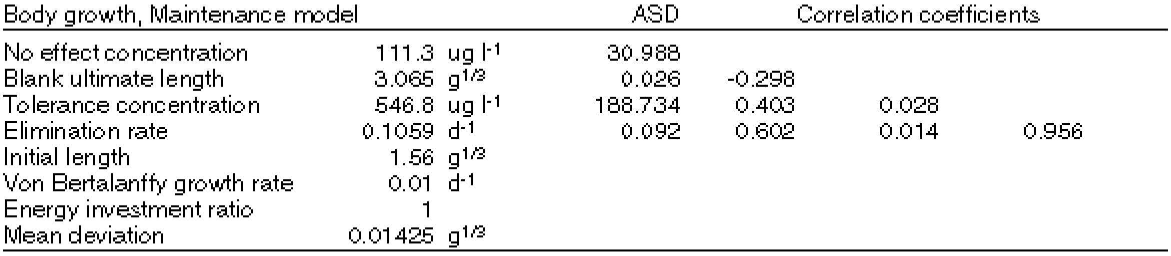


Figure 1: Effects of PCP on the survival of the fathead minnow *Pimephales promelas*. The profile ln likelihood function for the NEC, given in the graph on the left, has a non-typical shape, because DEBtox used few evaluation points. The reason is that the confidence interval of the NEC is here very small with respect to its value.



Body growth, Maintenance model			ASD	Correlation coefficients		
No effect concentration	111.3	ug l^{-1}	30.988			
Blank ultimate length	3.065	$g^{1/3}$	0.026	-0.298		
Tolerance concentration	546.8	ug l^{-1}	188.734	0.403	0.028	
Elimination rate	0.1059	d^{-1}	0.092	0.602	0.014	0.956
Initial length	1.56	$g^{1/3}$				
Von Bertalanffy growth rate	0.01	d^{-1}				
Energy investment ratio	1					
Mean deviation	0.01425	$g^{1/3}$				

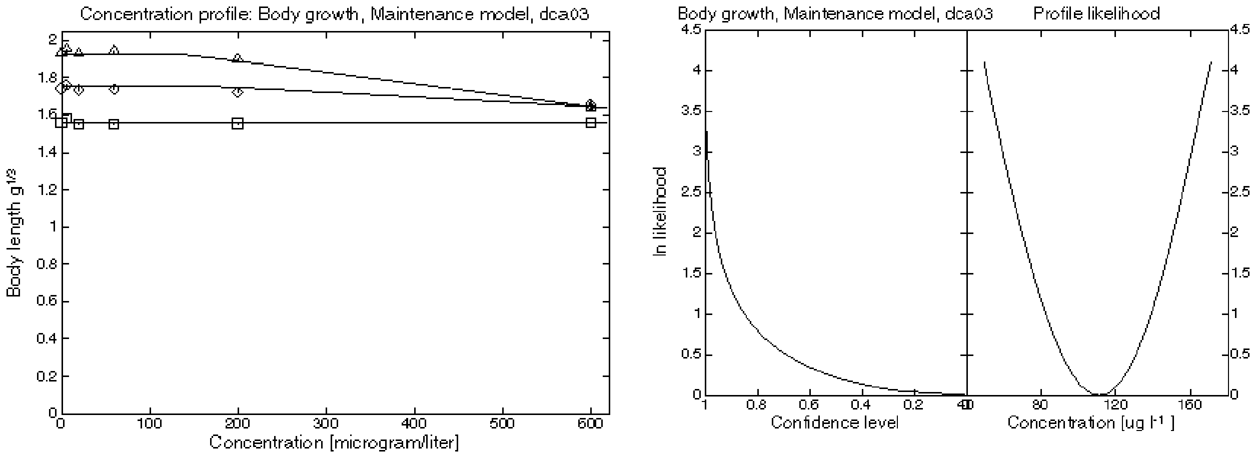
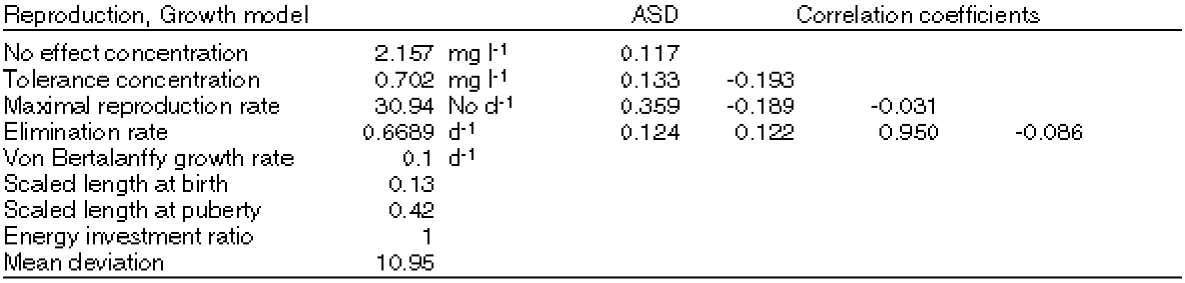


Figure: 2: Effects of DCA on the body growth of the rainbow trout *Oncorhynchus mykiss*. Data from the OECD ring test 1988/9. The costs for maintenance has been selected as target parameter. The profile ln likelihood function for the NEC, given in the graph on the left, is close to its large sample shape: the parabola.



Reproduction, Growth model			ASD	Correlation coefficients		
No effect concentration	2.157	mg l^{-1}	0.117			
Tolerance concentration	0.702	mg l^{-1}	0.133	-0.193		
Maximal reproduction rate	30.94	No d^{-1}	0.359	-0.189	-0.031	
Elimination rate	0.6689	d^{-1}	0.124	0.122	0.950	-0.086
Von Bertalanffy growth rate	0.1	d^{-1}				
Scaled length at birth	0.13					
Scaled length at puberty	0.42					
Energy investment ratio	1					
Mean deviation	10.95					

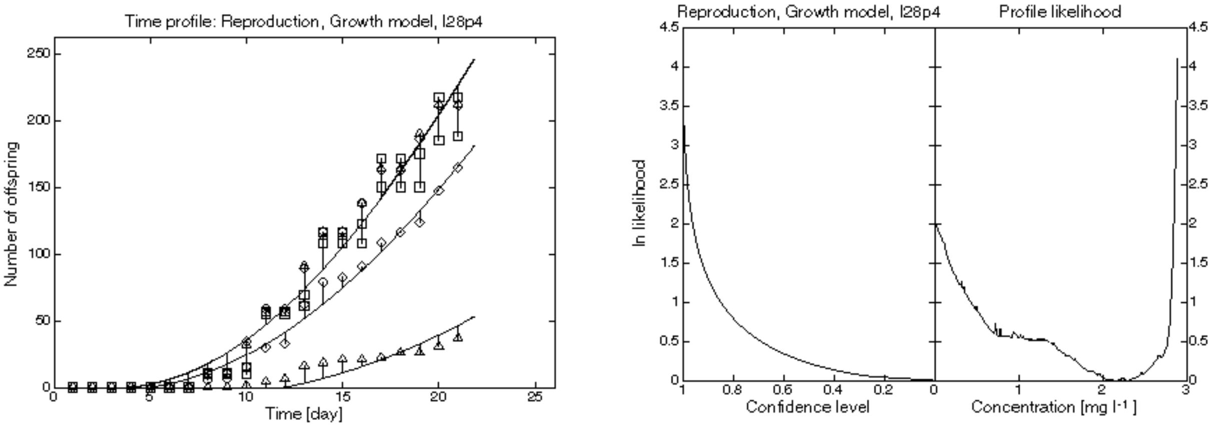
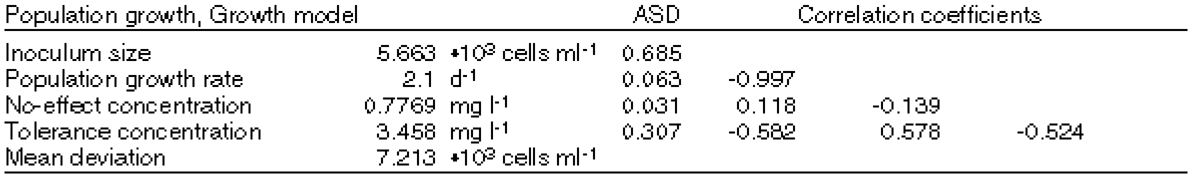


Figure 3: Effects of phenol on the reproduction of the water flea *Daphnia magna*. Data from the OECD ring test 1994/5. The costs for growth have been selected as target parameters. The little "teeth" in the profile ln likelihood function for the NEC, given in the graph on the left, are artifacts that resulted from numerical integrations to obtain the reproductive output. They are not typical; their occurrence depends on parameter values.



Population growth, Growth model			ASD	Correlation coefficients		
Inoculum size	5.663	$\cdot 10^3$ cells ml^{-1}	0.685			
Population growth rate	2.1	d^{-1}	0.063	-0.997		
No-effect concentration	0.7769	mg l^{-1}	0.031	0.118	-0.139	
Tolerance concentration	3.458	mg l^{-1}	0.307	-0.582	0.578	-0.524
Mean deviation	7.213	$\cdot 10^3$ cells ml^{-1}				

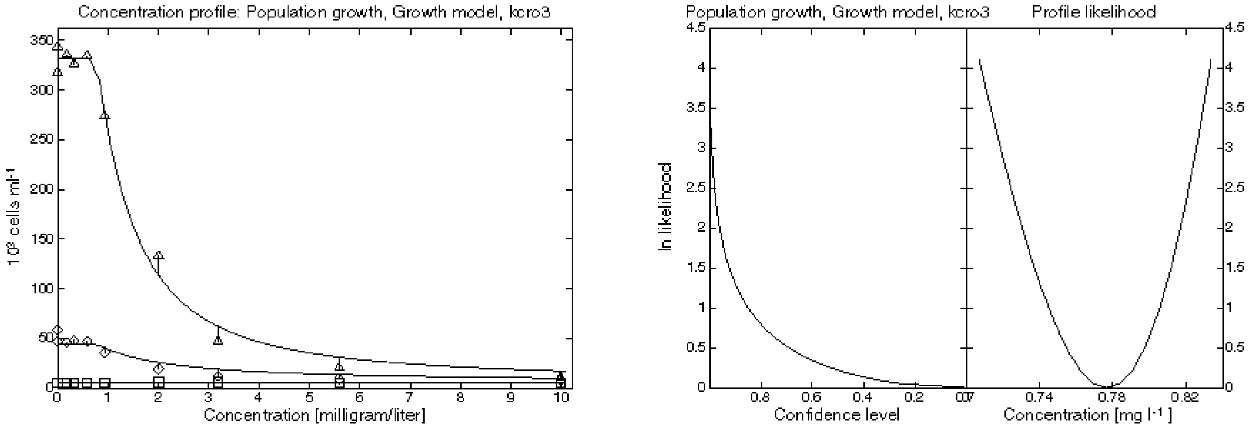


Figure 4: Effects of $K_2Cr_2O_7$ on the population growth rate of *Skeletonema costatum*. Data from the ISO ring test. The costs for growth has been selected as target parameter. The profile ln likelihood function for the NEC, given in the graph on the left, is close to its large sample shape: the parabola.

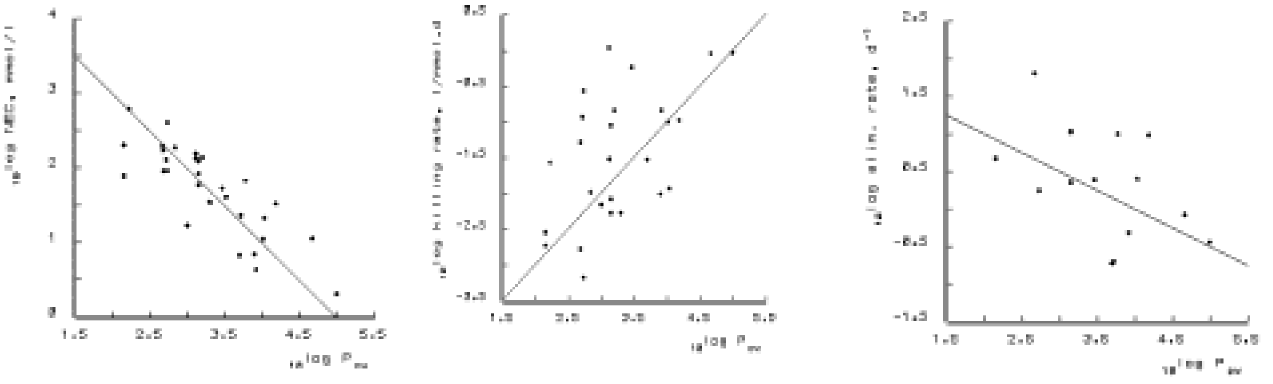


Figure 5: NEC, killing rate and elimination rate of alkyl benzenes as a function of the octanol/water partition coefficient. The nitro benzenes have been excluded. The slopes of the lines, *i.e.* -1, 1 and -0.5, respectively, follow from simple theoretical considerations. The data are from the 4d toxicity tests on survival of the fathead minnow, as presented in Geiger *et al* 1985-1990. The partition coefficients were obtained from Richardson & Gangolli or calculated according to Rekker 1977.

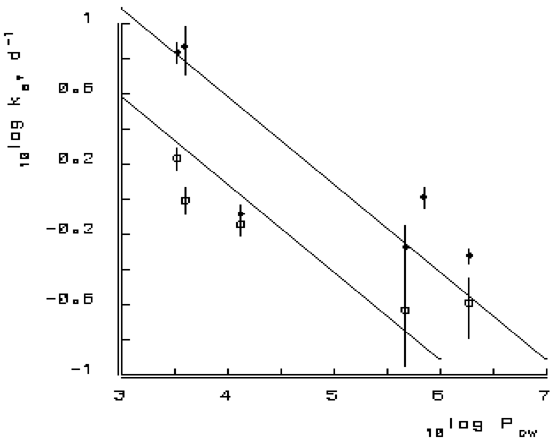


Figure 6: The elimination rate of alkyl benzenes as a function of the octanol/water partition coefficient for juvenile *Daphnia pulex* with a body length of 1mm (●) and for adult ones of 3mm (○). The lines correspond with the elimination rate = 392 / $\sqrt{P_{ow}}$ d^{-1} for juveniles and 124/ $\sqrt{P_{ow}}$ d^{-1} for adults. The ratio between these elimination rates, *i.e.* 3.17, corresponds well with the ratio of the body lengths, as expected from simple theoretical considerations. The data are from the 2d toxicity tests on survival, as presented in Hawker & Connell 1985.

4.1.2 Size of the organism

The EC50 and NOEC not only relate in a more complex way to the solubility in fat, compared to the dynamic parameters, but also to the body size of the test organism, by exactly the same argument. The larger the test organism, the longer it takes for effects to build up. This means that the LC50s for daphnia and fish are difficult to compare. The comparison of the parameters of the DEB-based model is easy, because the NEC and the killing rate (or tolerance concentration) do not depend on body size. The elimination rate is inversely proportional to a volumetric length, on the assumption that the exchange rate between organism and environment are proportional to the surface area of the organism. It is the only parameter that depends on body size. Figure 6 illustrates that these simple considerations make sense.

4.2 Comparisons on the basis of parameters

The logit model has three parameters: the blank response, the EC50 (LC50) and the slope. The NOEC is inconsistent with the logit model. Nonetheless it is frequently presented with the logit parameters in practice. It can be viewed as a fourth parameter which is 'estimated' in a rather odd way.

(It can take a very limited number of values and its estimation procedure has regrettable properties.) An extrapolation factor is used (except in population growth tests) to extrapolate to ultimate effects. This factor counts as a fixed parameter. The problem with this 'parameter' is that it relates to toxic effects, rather than to responses in the blank, which implies that it should not be the same for all chemicals.

The DEB-based model has four parameters; a blank response, a NEC, a tolerance concentration (or killing rate) and an elimination rate. In addition to this, for growth and reproduction tests only, it has three or four, respectively, parameters that are not estimated from the data. All these fixed parameters refer to details in the description of what happens in the blank. The blank response in the test does not provide the proper information for the fixed parameters. These parameters are

- scaled length at birth, *i.e.* the length at birth as a fraction of the maximum length in the blank, if the test would continue in time. This parameter applies to the toxicity test for reproduction. That for body growth uses a related parameter: the actual length at the start of the experiment. Although this fixed parameter can be treated as a free parameter, DEBtox treats it as a fixed one, because the size at the start of the experiment is frequently not measured in the case of early life stage tests.

- scaled length at puberty, *i.e.* the length at puberty as a fraction of the maximum length in the blank, if the test would continue in time. This parameter only applies to reproduction. Puberty is defined as the moment at which allocation to reproduction starts, which is somewhat earlier than the moment of first reproduction.

- energy investment ratio is a dimensionless parameter with a rather complex interpretation, which is described in the separate document. Realistic values for animals such as daphnia and fish are around 1. Numerical studies indicate that variations around this value have very little effect on the toxicity parameters.

- von Bertalanffy growth rate, with dimension time^{-1}. The value of this parameter is more important than of the previous one, but it can be measured easily and is known for hundreds of species.

The fixed parameters depend on the species, or even the strain, the culture conditions and details of the experimental protocol. If a particular laboratory has standardized its experimental protocol, these parameters need to be tuned only once, and not for each test. Part of the differences of toxicity results among laboratoria can be attributed to differences in fixed parameters. Good estimates can be given for the standard choices of species on the basis of existing ecophysiological data (Kooijman and Bedaux 1996).

The elimination rate in the DEB-based model stands for the rate at which effects build up during exposure. Its conceptual role is more or less comparable with the extrapolation factor in the static approach, but differs in a statistical sense. It is an ordinary model parameter, for which point-estimates as well as interval-estimates are available.

The slope parameter and the EC50 in the logistic model together play the same role as the tolerance concentration in the DEB-based model.

We can conclude that the DEB-based model has less parameters that are to be estimated from the results of the toxicity test than the logit model plus NOEC. On the basis of this measure for the complexity of a model, the simplest dynamic model is, therefore, simpler than the static one for all four toxicity tests.

4.3 Numerical comparisons

In this section we present numerical comparisons between the DEB-based model and the logit model plus the NOEC. Since the logit model only uses the response at the end of the experiment (apart from the algal growth inhibition test), we restrict the comparisons to EC*x* values for that exposure time. We compare the NOEC with the DEB-based EC0 for that exposure time, and not with the NEC, because the NEC relates to long-term exposure. The aim of this section is to compare statistics that are familiar to users of the static approach, not to show the potential of the dynamic one.

All calculations for the DEB-based model have been done with the software package DEBtox, as provided in Kooijman & Bedaux (1996). All calculations are based on nominal concentrations, except for the survival analyses, which are based on measured concentrations.

We fitted models for different modes of action of the compounds (*i.e.* target parameters) and noted that the differences in goodness of fit were generally small, and resulted in very similar values for the NEC.

Effects on survival

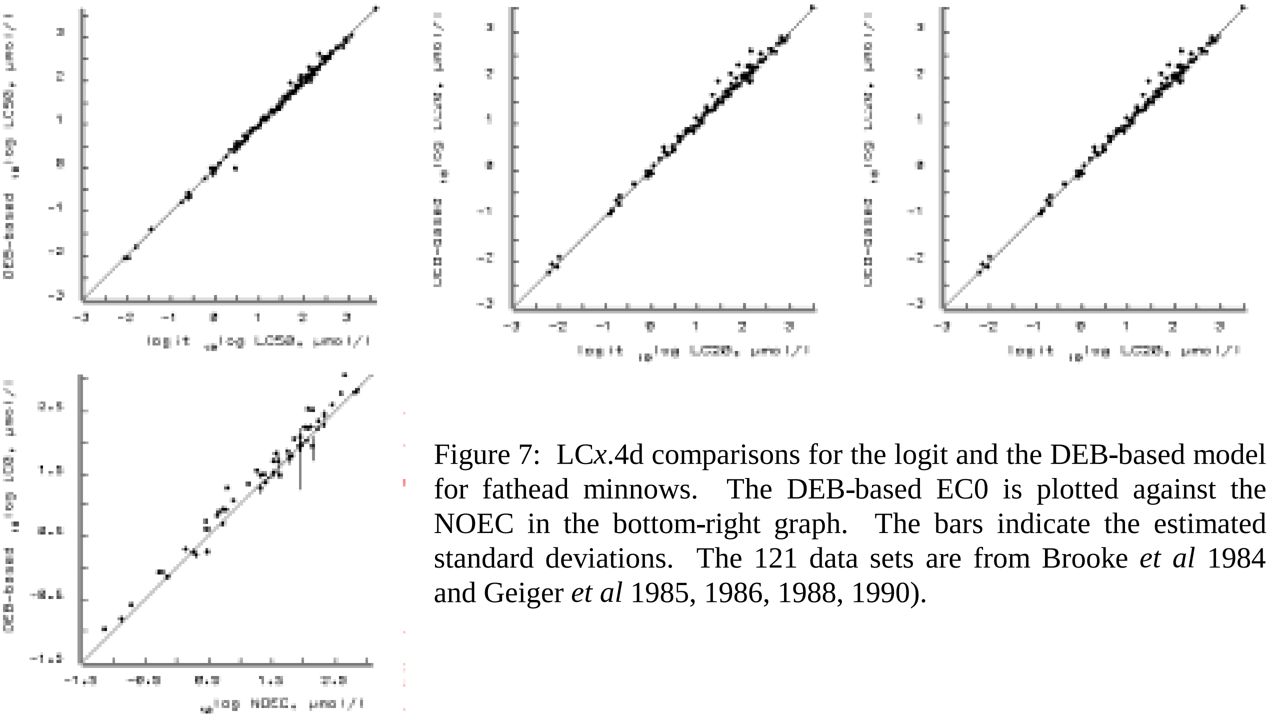
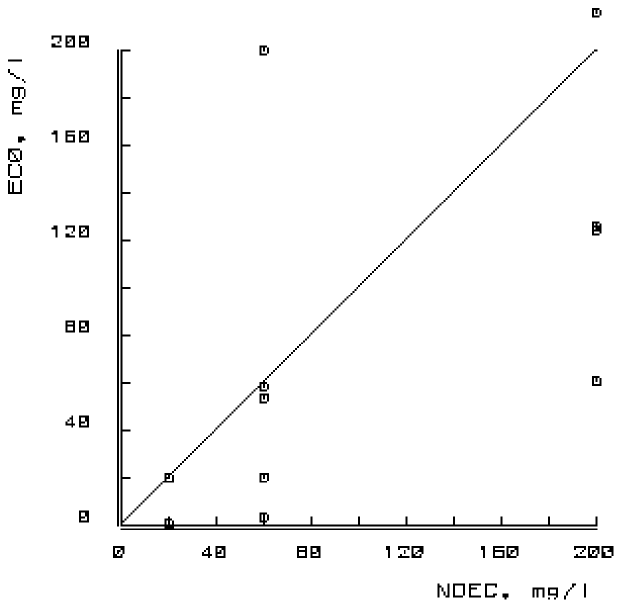


Figure 7: LC*x*.4d comparisons for the logit and the DEB-based model for fathead minnows. The DEB-based EC0 is plotted against the NOEC in the bottom-right graph. The bars indicate the estimated standard deviations. The 121 data sets are from Brooke *et al* 1984 and Geiger *et al* 1985, 1986, 1988, 1990).

4.3.1 Survival experiments

Figure 7 compares the LC*x*4d values for fathead minnows on the basis of the logit and the DEB-based model for many different chemicals. The logit parameters have been estimated according to the maximum likelihood method, as described in Kooijman 1981. We can conclude that the LC50 and LC20 values are very similar for both models, but the LC5 values for the DEB-based model tends to be higher than that of the logit model. This is to be expected, because LC*x* →0 for *x* →0 for the logit model, but LC0.4d ≥ NEC for the DEB-based model. The fact that the LC*x* values of the logit and DEB-based models correlate well is not surprising, because both models are fitted to the same data; very toxic chemicals result in low LC*x* values with both methods.



Effects on body growth

Figure 8: The EC0.28d for the DEB-based model for effects of DCA on the body growth of the rainbow trout is plotted against the NOEC on the basis of the Williams test for 11 toxicity tests of the OECD ringtest 1988. Effects on growth via effects on maintenance costs turned out to be the best fitting DEB-based model.

4.3.2 Body growth experiments

We used the data from the final ring test, organised by the OECD in 1988, with 3,4-dichloroaniline (DCA), for the rainbow trout *Oncorhynchus mykiss*. Since the volumetric lengths (*i.e.* the cubic root of weights) at the end of the toxicity tests differed little from those at the start, an analysis in terms of EC*x* values for the logit model is less appropriate: the EC50 would far exceed the highest tested concentration and the results would be most unreliable. We only present the results of the comparison of the NOEC with the EC0.28d in Figure 8.

The NOECs were identified on the basis of the test by Williams (Williams 1972), using a significance level of 5%. There were 16 individuals for each concentration of DCA.

The DEB-based model for effects on maintenance fitted best, although the differences in goodness of fit with the models for effects on growth and assimilation were usually small. The fixed parameters have been set at: initial length = mean initial length, energy investment ratio = 1, von Bertalanffy growth rate = 0.01 d^{-1}.

4.3.3. Reproduction experiments

We used the data from the final ring test, organised by the OECD in 1994/5, with 3,4-dichloroaniline (DCA), cadmium chloride and phenol. See Figures 9 and 10.

The NOEC values for these data were taken from Anonymous (1995), as listed for the case that the reproduction of females that die during the test are excluded.

The logit model has been fitted with SYSTAT version 5.02 using non-linear regression on the number of juveniles per adult for individuals that did not die in 21 days. The EC*x* values, and their standard deviations have been obtained via reparametrization of the logit model.

The data from DEBtox represent the cumulated total reproduction per living female, including all observation times. The reproduction of a female that died during the test has been included. No delay of reproduction could be observed for cadmium, and we selected the model for effects on the hazard of the ovum as the best fitting model for direct effects on reproduction. This model is rather similar to the other direct effect model, *i.e*. effect on costs per young. Where difference in goodness of fit for both models was relatively large (but still small absolutely), the hazard model fitted best. Some delay of reproduction could be observed for DCA and phenol, and we selected the model for effects on the growth costs as the best fitting model for indirect effects on reproduction. Since no data on the size of the adults are available, we could not test the predicted effects on growth. The fixed parameters have been set to the default settings of DEBtox: *i.e*. scaled length at birth = 0.13, scaled length at puberty = 0.42, energy investment ratio = 1, von Bertalanffy growth rate = 0.1 d^{-1}.

The growth model (or hazard model respectively) could be fitted to all data sets without problems, using DEBtox. The NECs and tolerance concentrations for data sets without effect of the toxicant were large with huge standard errors, as could be expected. We did not include data sets into the comparisons where the EC50.21d was much larger than the highest tested concentration.

Figures 9 and 10a-b represent the comparisons of the EC*x* values and the NOECs for the chemicals DCA, cadmium and phenol. The EC*x* values are quite comparable for both models if $x \geq 5$. The standard deviations of the EC*x* values of the logit model tend to be somewhat larger than for the DEB-based model. The EC1 values of the logit model tend to be lower than that of the DEB-based one. This can be expected, because the EC$x \rightarrow 0$ if $x \rightarrow 0$ for the logit model, while EC0 ≥ NEC for the DEB-based model. We also see that the scatter in the NOECs is larger than the scatter in the EC0s for two of the three chemicals.

Effects of DCA on Daphnia reproduction

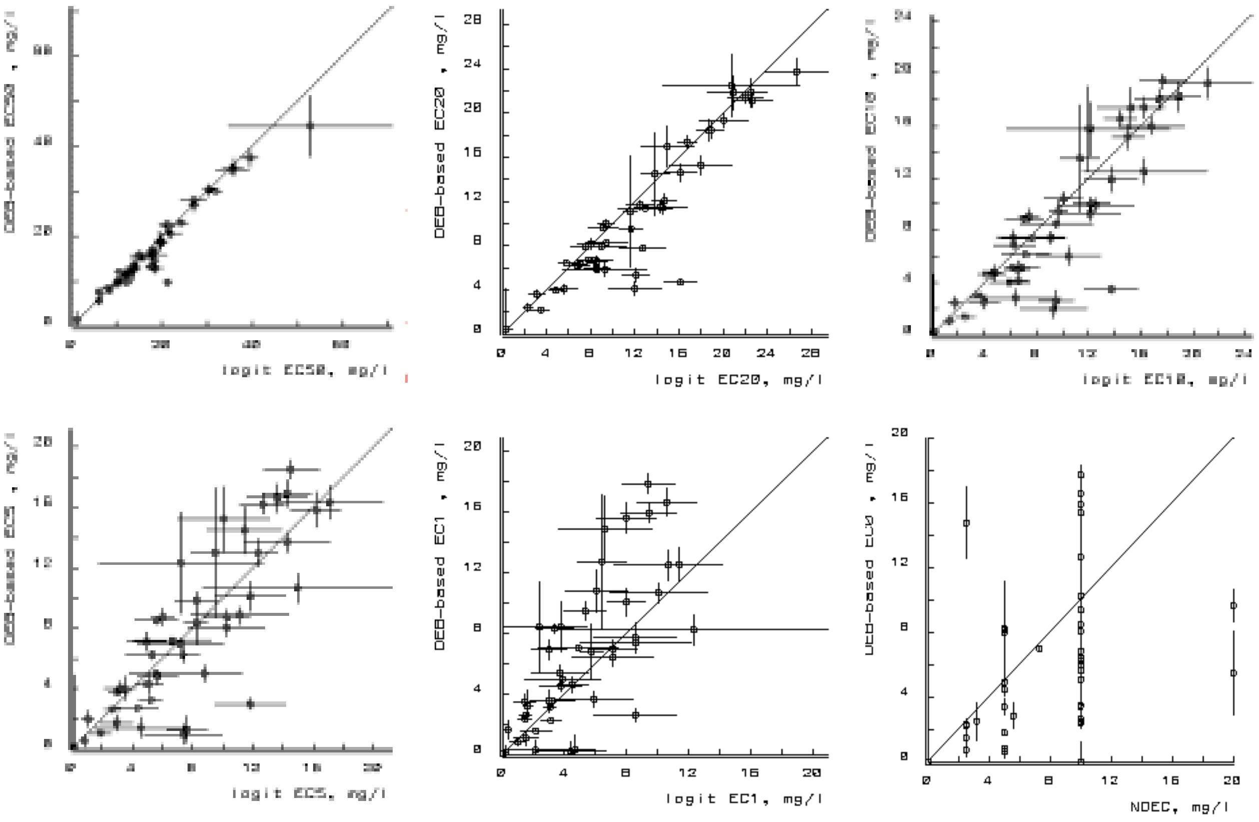


Figure 9: EC*x*.21d comparisons for the logit and the DEB-based model for the D*aphnia* reproduction test on DCA. The DEB-based EC0.21d is plotted against the NOEC. The bars indicate the estimated standard deviations. The data sets are from the final ring test of the OECD 1994/5. Four NOECs could not be obtained, and have been set to zero in the graph.

Effects of Cadmium on Daphnia Reproduction

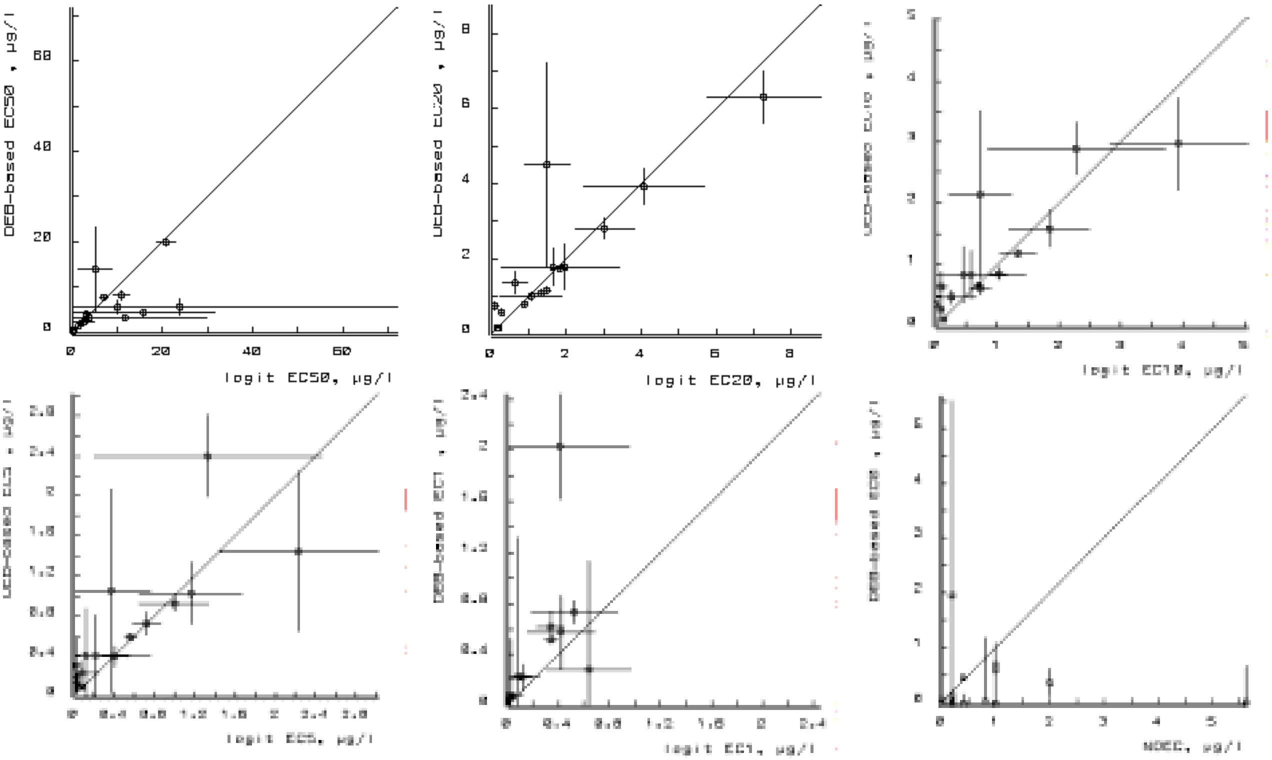


Figure 10a: EC*x*.21d comparisons for the logit and the DEB-based model for the *Daphnia* reproduction test on cadmium. The DEB-based ECO.21d is plotted against the NOEC. The bars indicate the estimated standard deviations. The data sets are from the final ring test of the OECD 1994/5.

Effects of phenol on Daphnia reproduction

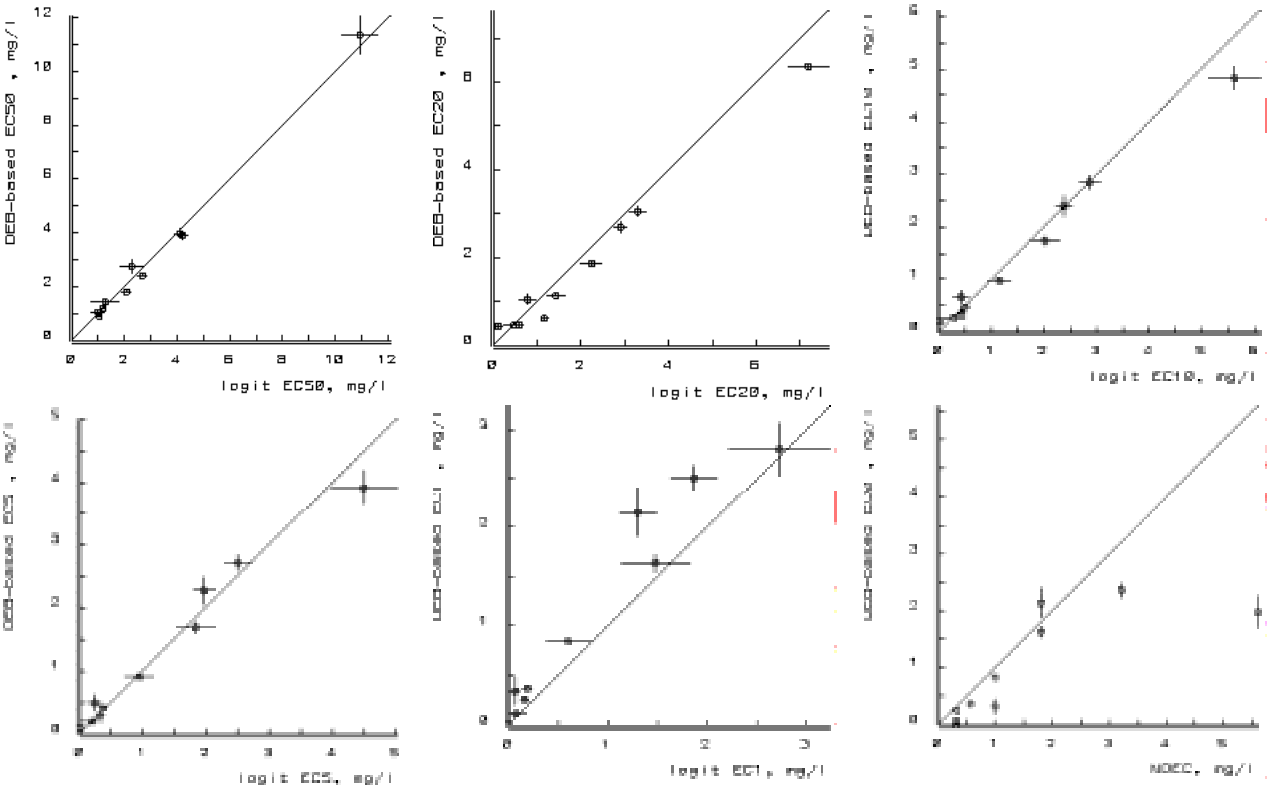


Figure 10b: EC*x*.21d comparisons for the logit and the DEB-based model for the *Daphnia* reproduction test on phenol. The DEB-based EC0.21d is plotted against the NOEC. The bars indicate the estimated standard deviations. The data sets are from the final ring test of the OECD 1994/5.

4.3.4 Population growth experiments

We analysed the data on the effects of 3,5 dichloro-phenol (DCP) and potassium dichromate from the ISO ring test with the diatoms *Skeletonema costatum* and *Phaeodactylum tricornutum* (marine algal growth inhibition test, ISO 10253, Geneva 1995).

The logit model was fitted to the data using non-linear regression, as described in Kooijman *et al* (1983), where the population growth rate decreases logistically with the concentration of test chemical. The standard deviations of the EC*x* values were obtained from those of the EC50, the slope and the covariance of both parameters, on the basis of the formulae presented in Kooijman and Bedaux (1996).

The DEB-based model for effects on growth costs fitted best for dichromate, while that for effects on the hazard rate fitted best for DCP. Since the DEB-based model assumes that toxicant kinetics is fast with respect to population growth, we have that EC0 = NEC.

Figures 11a-b compare the EC*x* values and the NOECs. The two data sets for dichromate, where the EC50 and the EC20 are small for the logit model and large for the DEB-based one, have already been indicated by Hanstveit (1991) as outliers. For DCP we have the opposite situation: two data sets for which the EC50 is large for the logit model and small for the DEB-based one. We see that the logit model has relatively small standard deviations for the EC50, but they rapidly become larger for the smaller effect levels to the extent that they become meaningless in quite a few data sets.

No proper NOEC could be identified in 4 of the 12 data sets for DCP and 6 of the 15 data sets for dichromate. These values have been set to zero in the graphs. The NEC was estimated to be zero in 1 data set for DCP and 1 for dichromate.

Effects of DCP on algal growth

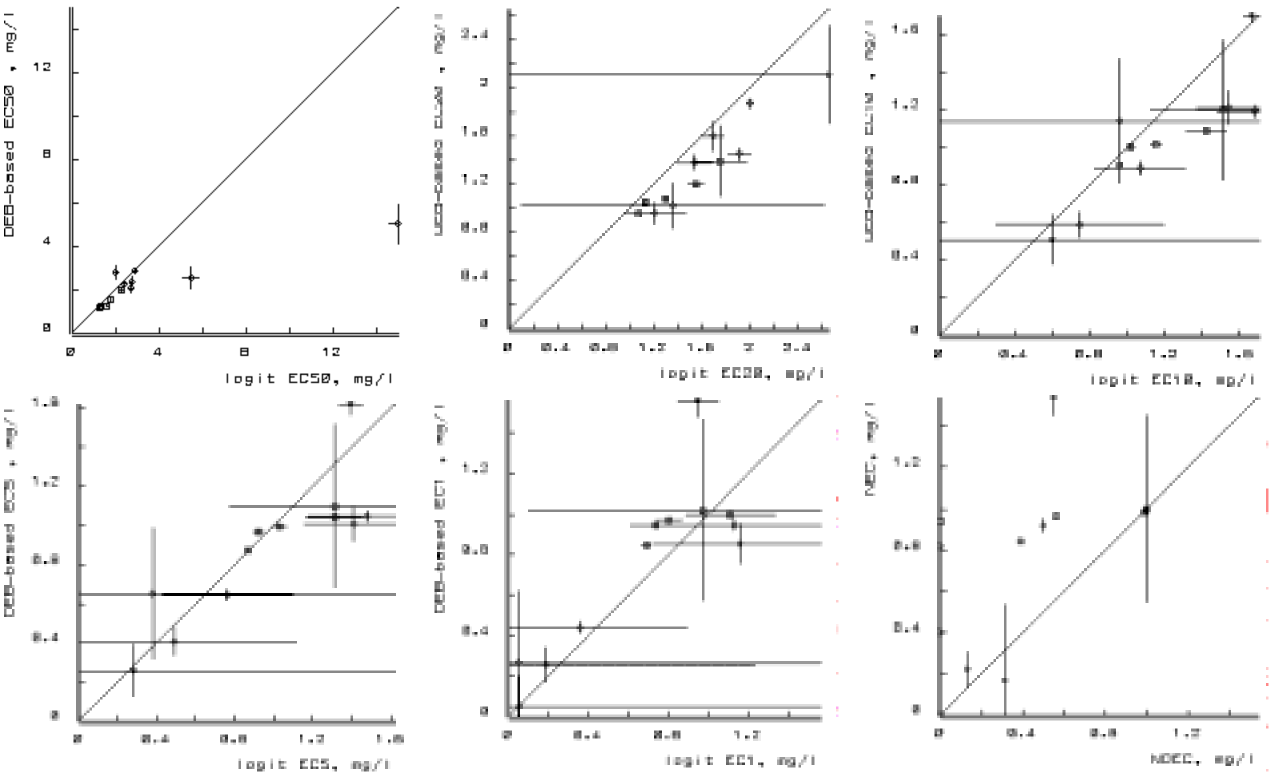


Figure 11a: EC*x* comparisons for the logit and the DEB-based model for the alga growth inhibition test on DCP. The DEB-based is plotted against the NOEC in the bottom-right graph. The bars indicate the estimated standard deviations. The data sets are from ISO ring tests.

Effects of Dichromate on algal growth

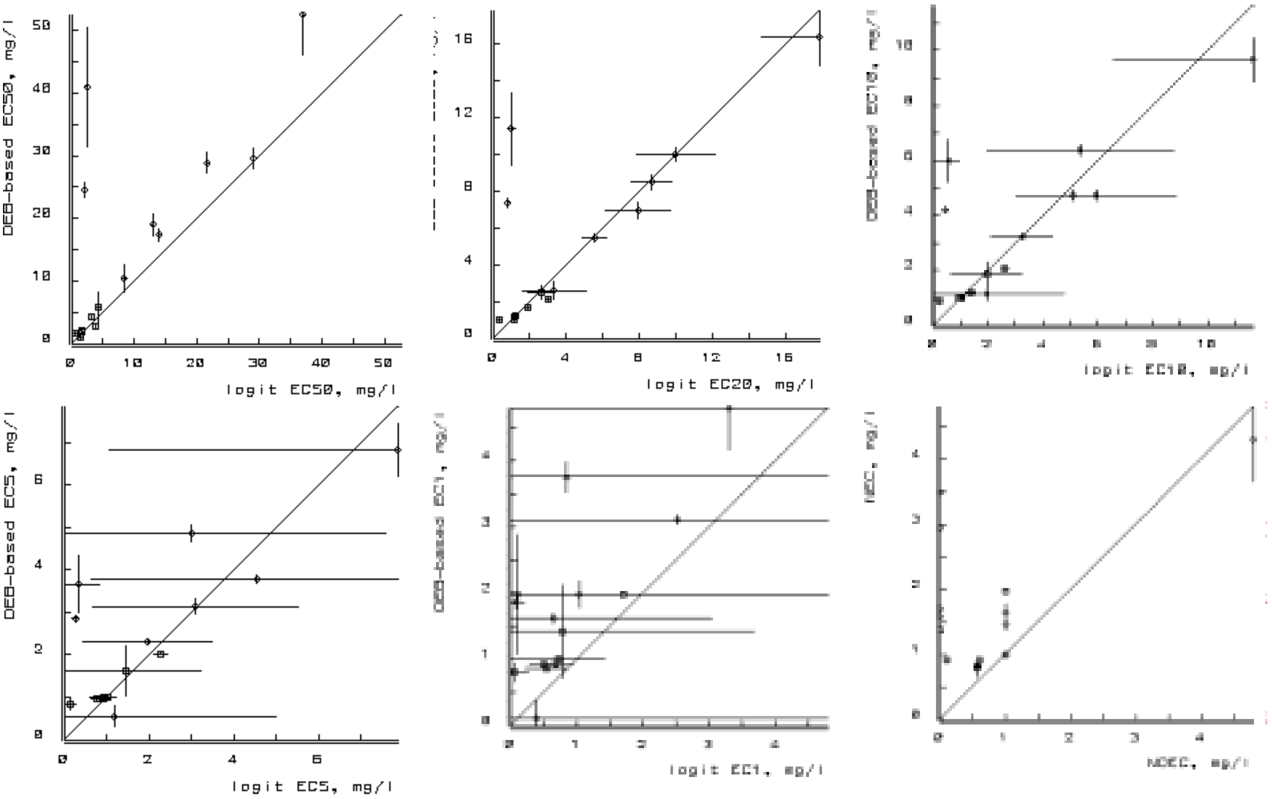


Figure 11b: EC*x* comparisons for the logit and the DEB-based model for the alga growth inhibition test on dichromate. The DEB-based is plotted against the NOEC in the bottom-right graph. The bars indicate the estimated standard deviations. The data sets are from ISO ring tests.

4.4 Advantages of the dynamic approach

- The DEB-based model allows the estimation of the NEC, which has good statistical properties, including an interval estimate.

- The dynamic approach needs no extrapolation factor to arrive at long-term effects. The elimination rate plays this role.

- The comparison of results of different toxicity tests is easy on the basis of the DEB-based model, because (i) its parameters are independent of the exposure time, (ii) it gives simpler QSARs and (iii) simpler relationships with the body size of the test animals.

- The static parameters (EC*x* values) can be obtained from dynamic ones, but not vice versa, which illustrates that the dynamic approach extracts more information from the same toxicity data.

- The dynamic approach uses all available data, while the static approach uses only the data at the end of the exposure. It has less parameters that are to be estimated from the data than the static approach.

- Additional information, such as the elimination rate, derived from toxico-kinetic data, can readily be used in the dynamic approach.

- The effects of repeated exposures and of pulse exposures can be evaluated readily, because of the dynamic properties of the DEB-based model.

- The toxicity of mixtures of compounds can readily be evaluated, including interactions between different compounds, due to the linearity of effect component of the DEB-based model.

- The effects of mutagenic compounds and ionogenic radiation can be quantified via effects on the target parameter ‘ageing acceleration’ (see separate document).

- Environmental risk assessment concerns effects of toxicants on ‘natural’ populations, not on individuals in a laboratory. To evaluate the consequences of toxic effects on individuals for population dynamics, we have to know how survival and reproduction change during the lifetime of an organism. This requires a dynamic approach.

- The DEB-based model is mechanistic, based on biological ideas. It allows a series of models from simple (routine testing) to complex (research) by changing the kinetics component.

4.5 Disadvantages of the dynamic approach

- The dynamic approach requires rather advanced numerical techniques for statistical evaluations. This problem is solved by the software package DEBtox.

- The blank component in the dynamic approach is more complex than in the static approach and involves biological knowledge. This knowledge also must be used in the application of the toxicity results in environmental risk assessment. Fixed parameters for the blank response reflect this biological realism. Supplementary eco-physiological data are required for each species of test organism to determine the values. These data are available for the species that are frequently used, and appropriate values for the fixed parameters are known.

- The different modes of action of the various compounds relate to different target parameters, which hampers the comparison of the toxicity of such compounds to some extent. The proper identification of the mode of action is not always feasible with the present experimental design. Additional observations, such as body length at the end of the *daphnia* reproduction experiment and/or the measurement of the feeding rate, would help substantially. Numerical results indicate that the proper identification of the mode of action is not essential for a reliable estimation of the NEC.

- The dynamic approach does make assumptions about the effects, assumptions that can be wrong. One should realize, however, that every approach is based on assumptions. The use of a fixed extrapolation factor to transform the estimated toxicity into a long-term toxicity in the static approach, for instance, assumes that this factor is the same for all compounds. The assumptions are more explicit and more realistic in the dynamic approach.

References

Anonymous 1995. Draft report of the final ring test of the *Daphnia magna* reproduction study. Draft OECD test guideline 202.

Brooke, L.T., Geiger, D.L., Call, D.J. & Northcott, C.E. 1984. *Acute toxicities of organic chemicals to fathead minnow* (*Pimephales promelas*) Vol **1**}. Center for Lake Superior Environmental Studies.
University of Wisconsin-Superior, USA.

Geiger, D.L., Brooke, L.T. & Call, D.J. 1990. *Acute toxicities of organic chemicals to fathead minnow* (*Pimephales* prom*e*las) Vol **5**. Center for Lake Superior Environmental Studies. University of Wisconsin-Superior, USA.

Geiger, D.L., Call, D.J. & Brooke, L.T. 1988. *Acute toxicities of organic chemicals to fathead minnow* (*Pimephales promelas*) Vol **2**. Center for Lake Superior Environmental Studies. University of Wisconsin-Superior, USA.

Geiger, D.L., Northcott, C.E., Call, D.J. \& Brooke, L.T. 1985. Acute toxicities of organic chemicals to fathead minnow Pimephales promelas Vol 5. Center for Lake Superior Environmental Studies. University of Wisconsin-Superior, USA.

Geiger, D.L., Poirier, S.H., Brooke, L.T. & Call, D.J. 1986. *Acute toxicities of organic chemicals to fathead minnow* (*Pimephales promelas*) Vol **3**. Center for Lake Superior Environmental Studies. University of Wisconsin-Superior, USA.

Hanstveit, A.O. 1991. The results of an international ring test of marine algal growth inhibition test according to ISO/DP 10253. TNO Report 91/236.

Hawker, D.W. & Connell, D.W. 1986. Bioconcentration of lipophilic compounds by some aquatic organisms. *Ecotox. Environ. Safety* **11**: 184-197.

Kooijman, S.A.L.M. 1981. Parametric analyses of mortality rates in bioassays. *Water Res*. **15**: 107-119.

Kooijman, S.A.L.M. & Bedaux, J.J.M. 1996. *The analysis of aquatic toxicity data*. VU University Press, Amsterdam (pp 160 + floppy)

Kooijman, S.A.L.M., Hanstveit, A.O. \& Oldersma, H. 1983. Parametric analyses of population growth in bioassays. *Water Res*. **17**: 727-738.

Rekker, R.F. 1977. The hydrophobic fragmental constant. Its derivation and application. A means of characterising membrane systems. In: Pharmacochemistry Library. Nauta, W.Th. & Rekker, R.F. (eds), Elsevier, Amsterdam.

Richardson, M.L. & Gangolli, S. 1995. *The dictionary of substances and their effects*. Publication of the Royal Society of Chemistry, Cambridge, CB4 4WF

Williams, D.A. 1972. The comparison of several dose levels with a zero dose control. *Biometrics* **28**: 519-531.

Appendix: Publications on DEB-based effect models

Individual-level

Bedaux, J.J.M. & Kooijman, S.A.L.M. 1994. Statistical analysis of bioassays, based on hazard modelling. *Envir. & Ecol. Stat.* **1**}: 303-314.

Haren, R.J.F.van, Schepers, H.E. Kooijman, S.A.L.M. 1994. Dynamic Energy Budgets affect kinetics of xenobiotics in the marine mussel *Mytilus edulis Chemosphere* **29**: 163-189

Kooijman, S.A.L.M. 1996. An alternative for NOEC exists, but the standard model has to be replaced first. *Oikos* **75**: 310-316.

Kooijman, S.A.L.M. 1996. Process-oriented descriptions of toxic effects. In: Schüürmann, G. and Markert, B. (eds) *Ecotoxicology*, John Wiley. (to appear)

Kooijman, S.A.L.M. & Bedaux, J.J.M. 1996. Some statistical properties of estimates of no-effects concentrations. *Water Res.* **30**: 1111-1111.

Kooijman, S.A.L.M. & Bedaux, J.J.M. 1996. Analysis of toxicity tests on *Daphnia* survival and reproduction. *Water Res.* **30**: 1711-1723.

Kooijman, S.A.L.M. & Bedaux, J.J.M. 1996. Analysis of toxicity tests on fish growth. *Water Res.* **30**: 1633-1644.

Kooijman, S.A.L.M. & Bedaux, J.J.M. 1996. The analysis of aquatic toxicity data. UV University Press, Amsterdam (pp 160 + floppy)

Kooijman, S.A.L.M., Bedaux, J.J.M. & Slob, W. 1996. No-Effect Concentration as a basis for risk assessment. *Risk Analysis* **16** (4) (to appear)

Kooijman, S.A.L.M., Hanstveit, A.O. & Nyholm, N. 1996. No-effect concentrations in alga growth inhibition tests. *Water Res.* **30**: 1625-1632.

Kooijman, S.A.L.M. & Haren, R.J.F.van 1990. Animal energy budgets affect the kinetics of xenobiotics. *Chemosphere* **21**: 681-693.

Population level

Hallam, T.G., Lassiter, R.R. & Kooijman, S.A.L.M. 1989. Effects of toxicants on aquatic populations. In: Levin, S.A., Hallam, T.G. & Gross, L.F. (eds) *Mathematical Ecology*,. Springer, London pp 352-382.

Kooijman, S.A.L.M. 1991. Oecotoxicologische risico-evaluatie. In: Straalen, N.M. van & Verkleij, J.A.C. (eds) *Leerboek Oecotoxicologie*: 348-356. VU University Press, Amsterdam.

Kooijman, S.A.L.M. 1991. Effects of feeding conditions on toxicity for the purpose of extrapolation. *Comp. Biochem. Physiol.* **100c(1/2)**: 305-310.

Kooijman, S.A.L.M. 1985. Toxicity at population level. In: Cairns, J. (ed). *Multispecies toxicity* testing: 143-164. Pergamon Press, N.Y.

Kooijman, S.A.L.M. 1988. Strategies in ecotoxicological research. *Environmental Aspects of Applied Biology* **17** (1): 11-17.

Kooijman, S.A.L.M., Hanstveit, A.O. & Hoeven, N. van der 1987. Research on the physiological basis of population dynamics in relation to ecotoxicology. *Wat. Sci. Tech.* **19**: 21-37.

Kooijman, S.A.L.M. & Metz, J.A.J. 1983. On the dynamics of chemically stressed populations; the deduction of population consequences from effects on individuals. *Ecotox. Envir. Safety* **8***:* 254-274.

Straalen, N.M. van, Kooijman, S.A.L.M., Eijsackers, H.J.P. & van Leeuwen, C.J. 1988. Behoefte aan ecotoxicologische modellen groeit. *Chemisch Magazine maart* 1988: 186-193.

ANNEX 9

The Dynamic Energy Budget (DEB) model

S.A.L.M. Kooijman, 12 Sept 96

This short note introduces the Dynamic Energy Budget (DEB) model which specifies the rules for uptake and use of food for ectothermic ('cold-blooded') animals. The term 'Dynamic' refers to the change of the energy budget during the life history of an animal, see Figure 1. Three stages are distinguished: the embryo (which does not eat), the juvenile (which eats, but does not reproduce) and the adult (which eats and reproduces). With minor modifications, the model also applies to endothermic (``warm-blooded") animals and unicellulars (including bacteria) that are limited in growth by a single resource.

The diagram in Figure 2 presents the fluxes of energy through an animal, as conceived in the DEB model. Energy is extracted from food and added to the reserves, *i.e.* a combination of carbohydrates, lipids and proteins. Energy in the reserves is used for four destinations, which can be combined into two groups of two: growth (*i.e.* increase in structural biomass, mainly in the form of proteins) plus somatic maintenance (including activity, protein turnover, *etc.*) and maturation (*i.e.* development, the increase in the state of maturity) plus maturity maintenance (*i.e.* maintaining the acquired state of maturity). Adults do not longer invest into maturation, but into reproduction. The various destinations only compete within each group. So, the animal ceases growth when the energy allocated to growth plus somatic maintenance is fully required for somatic maintenance. Under these conditions, it can continue to reproduce (if it is an adult), because reproduction is not in the same group of destinations. Likewise, reproduction ceases when the energy allocation to reproduction plus maturity maintenance is fully required for maturity maintenance.

The rules, presented as axioms in Table 1 quantify the fluxes that are shown in Figure 2. Each of these axioms can be justified mechanistically, and has been tested against experimental data for a wide variety of species. These simple rules have a myriad of implications for suborganismal organisation and population dynamics. For instance, the energy costs of an egg and its incubation time follow directly from these rules. Although the rules define energy fluxes, all mass fluxes, including respiration (*i.e.* oxygen use or carbon dioxide production) and the rate of nitrogen waste

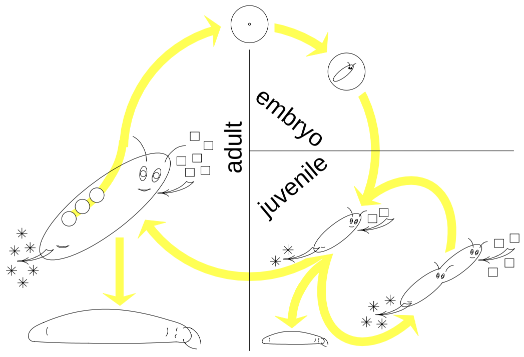


Figure 1: Dynamic Energy Budget theory quantifies the energetics as it changes during life history. The key processes are feeding, digestion, storage, maintenance, growth, development, reproduction and ageing. Dividing organisms, such as microbes, are included by conceiving them as juveniles.

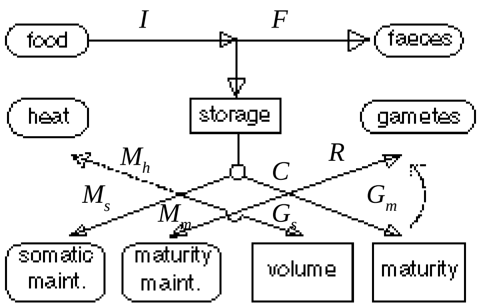


Figure 2: Energy fluxes through an animal: *I* ingestion (uptake), *F* defecation, *A* assimilation, *C* catabolic, M_s somatic maintenance, M_m maturity maintenance, M_h heating (endotherms), G_s somatic growth, G_m maturation, *R* reproduction. The rounded boxes indicate sources or sinks. All fluxes contribute a bit to dissipating heat, but this is not indicated in order to simplify the diagram.

(ammonia, urine) also follow from these rules, via the conservation law for mass. The rules give an explanation for the observed increase in respiration coupled to the feeding process (this previously poorly understood phenomenon is called the ‘specific dynamic action’). It can be shown that the rules provide a theoretical basis for the widely applied method of indirect calorimetry, where measurements of oxygen use, carbon dioxide production and nitrogen waste are used to obtain the flux of dissipating heat.

The DEB model specifies the uptake and use of food by an animal as a dynamic system with three state variables (volume of structural biomass, amount of reserves and cumulated damage) and 11 parameters:

L_b	length at birth	L_p	length at puberty	K	saturation constant
$\{I_m\}$	max. spec. ingestion rate	$\{A_m\}$	max. spec. assim. rate	P_a	ageing acceleration
[M]	spec. somatic maint. costs	[G]	spec. growth costs	$[E_m]$	max. spec. reserves
κ	$\frac{\text{somatic maint. + growth costs}}{\text{catabolic energy}}$	q	overhead costs of reprod.		

Although the number of parameters might seem large, it is in fact extremely small in view of the number of processes hat are specified. Only a small selection of these parameters is involved in the description of any particular measured variable. If we evaluate the expression for size as a function of age, for instance, we know beforehand that parameters with energy in their dimensions will occur only as ratios, such that the dimension energy drops out. This is because energy is not involved directly in size measurements. We need to know the value of a parameter that has energy in its dimensions only if we want to describe energies.

Three compound parameters frequently appear in expressions for physiological quantities (cf Figure 3): the maintenance rate constant, m =[M]/[G] (dimension: time^{-1});the energy conductance, v = {A_m}/[E_m] (dimension: length time^{-1}); the energy investment ratio, g = [G]/κ[E$_m$] (dimension: none).

Figure 3 presents the feeding-at-length, respiration-at-length, growth-at-age and reproduction-at-age of the waterflea *Daphnia magna* for the situation of constant food density and temperature. This species, like most other species of animal, hardly changes its shape during growth, which implies that its surface area is proportional to the squared volumetric length and its volume to the cubed length. The four relationships in Figure 4 cover the major processes of uptake and use of food. The expressions, which follow from the set of rules of Table 1, show how the (compound) parameters in the description of these relationships depend on the feeding conditions. The scaled functional response *f* (defined as the ratio of the ingestion rate and the maximum one for an animal of that size) is under experimental control. Length-at-age turns out to follow the von Bertalanffy growth curve. By choosing different feeding levels, the von Bertalanffy growth rate (which is a compound parameter) and the ultimate length (another compound parameter) change in a particular way. This information can be used to estimate the compound parameters (*m, v* and *g)* that are involved. These compound parameters can also be estimated from data such as the specific rate of weight decrease during starvation, respiration ontogeny during the embryonic period and survival probability-at-age.

The most far reaching and spectacular implications of the rules are the inter-specific body-size-scaling relationships. These relationships give trends in parameter values as they covary over different species (bacteria to whales). The 11 parameters can be classified in two groups. One group of parameters does not depend on body size, while the other group does depend on body size in a simple and predictable way: these parameters are proportional to the volumetric length, *i.e*. the cubic root of the body volume. Deviations of parameter values from these trends reflect ecophysiological adaptations of that species. All physiological variables that can be written as functions of the parameters can, for this reason, also be written as functions of (maximum) body size. Many of these functions have been worked out, tested against data and found to be realistic. Among them is the respiration rate, which turns out to be a weighted sum of squared and cubic (volumetric) length. This is very similar to empirical relationships, that indicate that respiration is approximately proportional to body mass to the power 0.75. The DEB model solves the long standing problem of understanding this empirical relationship.

The intra-specific body-size-scaling relationships (where we have just one set of parameters to describe the processes of food uptake and use) are fundamentally different from inter-specific body-size-scaling relationships (where we have ‘sloppy’ trends in parameter values among species). Hence, the fact that respiration, as it increases during the growth of an individual, turns out to be a weighted sum of squared and cubic length, just like for inter-specific comparisons, is merely coincidence. The volume-specific respiration decreases with body volume during life, because of the decreasing investment into growth.

A full description of the theory can be found in: Kooijman, S.A.L.M. 1993. *Dynamic Energy Budgets in Biological Systems*. *Theory and applications in ecotoxicology*. Cambridge University Press, ISBN 0-521-45223-6, 350 pp.

Table 1: Key assumptions of the DEB model that specify a dynamical system with the state variables structural body mass, reserves and cumulated damage.

- If the investment into maturation exceeds a given threshold value, the organism changes its stage, *i.e*. it switches from the embryonic stage to the juvenile stage by initiating the feeding process, or from the juvenile stage to the adult stage by ceasing maturation and initiating the production of gametes (eggs, sperm).

- Food uptake is proportional to surface area and depends hyperbolically on food density *(i.e*. Holling type II functional response).

- The specific reserve dynamics is first order, with a rate that is inversely proportional to the volumetric length.

- The allocation to somatic maintenance plus growth (*i.e*. increase in structural biomass) is a fixed fraction of the energy drain from the reserves, which further includes maturity maintenance plus maturation or reproduction. This rule is called the κ rule.

- Homeostasis of structural biomass and reserves, *i.e*. their chemical composition does not change, despite changes in the chemical composition of the environment. Since the amount of reserves can change relative to the structural biomass, certain changes in the chemical composition of the individual as a whole are possible. The homeostasis assumption implies that the following items are constant.

  food-energy conversion, although it depends on the type of food;
  volume-specific maintenance costs (both somatic and maturity);
  volume-specific growth costs.

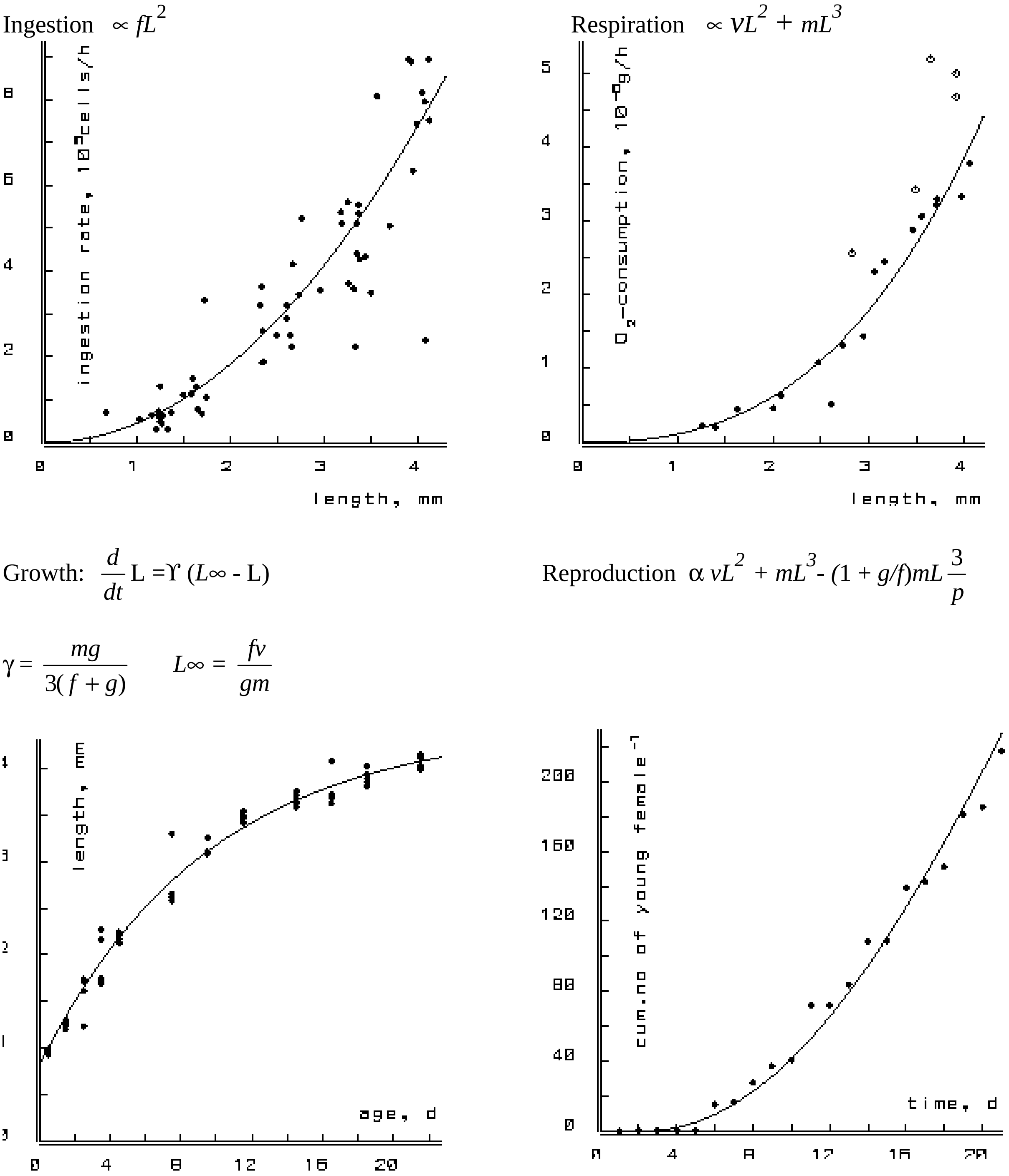
- The hazard rate is proportional to the accumulated damage.

  the damage production is proportional to the changed DNA;
  the DNA change is proportional to the use of oxygen.

- The initial conditions are given by

  the initial structural biomass is negligibly small;
  the reserve density at birth equals that of mother at egg laying;
  the initial damage is negligibly small.

Figure 3: Investment into maturation and the κ-rule for allocation of energy from reserves solves the following puzzle: *Daphnia magna* starts to reproduce upon exceeding 2.5 mm body length. Reproduction takes about 80% of the budget. Where does this energy come from? Ingestion or respiration is not rapidly increased at this size, nor is growth reduced. The rules in Table 1 imply the expressions presented above the graphs, where *L* stands for body length, *a* for age, and the compound parameters are given in the text.



Ingestion $\propto fL^2$

Respiration $\propto vL^2 + mL^3$

Growth: $\dfrac{d}{dt}\text{L} = \Upsilon\,(L\infty - \text{L})$

Reproduction $\alpha\ vL^2 + mL^3 - (1 + g/f)mL\dfrac{3}{p}$

$$\gamma = \frac{mg}{3(f+g)} \qquad L\infty = \frac{fv}{gm}$$

Appendix: Publications on the DEB-model

General and individual-level

Evers, E.G. & Kooijman, S.A.L.M. 1989. Feeding and oxygen consumption in *Daphnia magna*; A study in energy budgets. *Neth. J. Zool*. **39**: 56-78.

Haren, R.J.F. 1995. Application of Dynamic Energy Budgets to xenobiotic kinetics in *Mytilus edulis* and population dynamics of *Globodera pallida*. PhD-thesis, Vrije Universiteit, June 1995.

Haren, R.J.F. van & Kooijman, S.A.L.M. 1993. Application of the dynamic energy budget model to *Mytilus edulis* (L). *Neth. J. Sea Res* **31**: 119-133.

Konarzewski, M. & Kooijman, S.A.L.M. 1994. Models for growth in precocial and altricial birds. *J. Ornith*. **135** (3): 324.

Konarzewski, M., Kooijman, S.A.L.M. & Ricklefs, R.E. 1996. Models for avian growth and development. In: Starck, J.M. & Ricklefs, R.E. (eds) *Avian growth and development*. Oxford University Press (to appear).

Kooijman, S.A.L.M. 1986. Energy budgets can explain body size relations. *J. Theor. Biol*. **121**: 269-282.

Kooijman, S.A.L.M. 1986. What the hen can tell about her egg; Egg development on the basis of budgets. *J. Math. Biol*. **23**: 163-185.

Kooijman, S.A.L.M. 1988. The von Bertalanffy growth rate as a function of physiological parameters; A comparative analysis. In: Hallam, T.G., Gross, L.J., & Levin, S.A. (eds) *Mathematical ecology*: 3-45. World Scientific, Singapore.

Kooijman, S.A.L.M. 1993. *Dynamic Energy Budgets in Biological Systems. Theory and applications in ecotoxicology*. Cambridge University Press, 350 pp.

Kooijman, S.A.L.M. 1994. Effects of temperature on birds. In: Hagemeijer, E.J.M. & Verstrael, T.J. *Bird Numbers* 1992. *Distribution, monitoring and ecological aspects*. 12th Internat. Conf. of IBCC and EOAC., 14-18 Sept 92, Noordwijkerhout, pp 285-290.

Kooijman, S.A.L.M. 1995. The stoichiometry of animal energetics. *J. Theor. Biol*. **177**: 139-149.

Ratsak, C.H., Kooijman, S.A.L.M. & Kooi, B.W. 1993. Modelling of growth of an oligochaete on activated sludge. *Water Research* **27**: 739-747.

Ratsak, C.H., Kooi, B.W. & Kooijman, S.A.L.M. 1995. Modelling the individual growth of *Tetrahymena sp*. and its population consequences. *J. Euk. Microbiol*. **42**: 268-276.

Stouthamer, A.H. & Kooijman, S.A.L.M. 1993 Why it pays for bacteria to delete disused DNA and to maintain megaplasmids. *A. van Leeuwenhoek* **63**: 39-43.

Visser, J.A.G.M. de, Maat, A. ter & Zonneveld, C. 1994 Energy budgets and reproductive allocation in the simultaneous hermaphrodite pond snail, *Lymnaea stagnalis* (L.): A trade-off between male and female function. Am. Nat. **144**: 861-867.

Zonneveld, C. & Kooijman, S.A.L.M. 1989. The application of a dynamic energy budget model to *Lymnaea stagnalis*. *Functional Ecology* 3: 269-278.

Zonneveld, C. & Kooijman, S.A.L.M. 1993. Comparative kinetics of embryo development. *Bull. Math.* Bi*o*l. **3**: 609-635.

Zonneveld, C. & Kooijman, S.A.L.M. 1993. Body temperature affects the shape of avian growth curves. *J. Biol. Syst.* **1**: 363-374.

Zonneveld, C. & Kooijman, S.A.L.M. 1989. Application of a general energy budget model to *Lymnaea stagnalis. Functional Ecology* **3**: 269-278.

Zonneveld, C. 1992. Animal energy budgets: a dynamic approach. PhD-thesis, Vrije Universiteit Dec 92.

Population level

Kooijman, S.A.L.M. 1986. Population dynamics on the basis of budgets. In: Metz, J.A.J. & Diekmann, O. (eds). *The dynamics of physiologically structured populations*: 266-297. Springer Lecture Notes in Biomathematics. Springer-Verlag, Berlin.

Kooijman, S.A.L.M. 1992. Biomass conversion at population level. In: DeAngelis, D.L. & Gross, L.J. (eds) *Individual based models; an approach to populations and communities.:* 338-358. Chapman & Hall.

Kooijman, S.A.L.M. 1994. Individual based population modelling. In: Grasman, J. & Straten, G. van (eds): *Predictability and Nonlinear Modelling in Natural Sciences and Economics*. Kluwer Academic Publishers, Dordrecht: 232-247.

Kooijman, S.A.L.M., Hoeven, N. van der & Werf, D.C. van der 1989. Population consequences of a physiological model for individuals. *Functional* Ecolo*g*y **3**: 325-336.

Kooijman, S.A.L.M. & Kooi, B.W. 1995. Catastrophic behaviour of myxamoebae. *Nonlin. World* **3**: 77-83.

Kooijman, S.A.L.M., Kooi, B.W. & Boer, M.P. 1995. Rotifers do it with delay. The behaviour of reproducers vs dividers in chemostats. *Nonlin. World* **3**: 107-128.

Kooijman, S.A.L.M., Muller, E.B. & Stouthamer, A.H. 1991. Microbial dynamics on the basis of individual budgets. *Antonie van Leeuwenhoek* **60**: 159-174.

Kooi, B.W. & Boer, M.P. 1995. Discrete and continuous time population models, a comparison concerning proliferation by fission. *J. Biol. Systems* **3**: 543-558.

Kooi, B.W. & Kooijman, S.A.L.M. 1994. Existence and stability of microbial prey-predator systems. *J. Theor. Biol* **170**: 75-85.

Kooi, B.W. & Kooijman, S.A.L.M. 1994. The transient behaviour of food chains in chemostats. *J. Theor. Biol* **170**: 87-94.

Kooi, B.W. & Kooijman, S.A.L.M. 1995. Many limiting behaviours in microbial food chains. In: Arino, O., Kimmel, M. & Axelrod, D. (eds) *Mathematical Population Dynamics*., vol 2 Wuerz Publ, Winnipeg, Canada: 131-148.

Kooi, B.W. & Kooijman, S.A.L.M. 1995. Mass balance versus logistic equation in food chains. Proc. 4th Internat. Conf. Math. Pop. Dynamics, Houston. (to appear).

Kooi, B.W., Boer, M.P. & Kooijman, S.A.L.M. 1995. On the use of the logistic equation in food chains. (*Ecol. Mod*., subm.).

ANNEX 10

A review of statistical data analysis and experimental design in OECD aquatic toxicology Test Guidelines

Simon Pack

Shell Research Ltd.
Sittingbourne Research Centre
Sittingbourne, Kent, ME9 8AG, U.K.

August 1993

Summary

OECD aquatic toxicology Test Guidelines are reviewed with regard to the advice given on experimental design and data analysis. The aims are:

(1) to review existing Guidelines with a view to improving the test design and data analysis in individual Guidelines where necessary.

(2) to harmonise data analysis, as far as possible, across Guidelines.

(3) to make general recommendations about the use of statistics, taking into account recent trends in the scientific literature.

It is concluded that experimental design recommendations may be too restrictive in some circumstances. Furthermore, it is felt that too little help and guidance is given on the data analysis.

There is debate in the scientific literature about the use of the no-observed-effect-concentration (NOEC). After considering the issues it is concluded that the NOEC is not the preferred summary measure of toxic effect. It is recommended that EC point estimation (*i.e.* percentiles of the concentration-response curve) should become the preferred option.

The design and analysis of an experiment are closely linked. A move away from NOECs to EC point estimation may lead to a reduction in the costs of experiments in situations where the experimenter has some reasonable prior estimate of the concentration-response curve. This is because such information can be used to derive designs that may involve fewer concentrations or employ less replication than presently recommended in the Guidelines.

The main recommendations of this report are:

(1) That EC point estimation should be considered as the preferred type of analysis.

(2) That Member countries should consider extending their network of experts to include a list of statistical experts to assist in the development of current and future Guidelines.

(3) That OECD should promote the development of a handbook of statistical methods and computer software to assist in the test design and data analysis for the aquatic toxicology Guidelines.

Contents

1. Introduction

There is an increasing awareness in society of the need to protect the environment. The role of science is to provide decision makers with reliable information on the impact of man's activities.

The aquatic environment is one of the major areas of concern since many chemicals and waste materials may ultimately find their way into water courses. Furthermore, aquatic life is often very sensitive to pollution and damage here can have important consequences for major food chains.

To provide a framework for assessing the toxic effects of chemicals on aquatic organisms, national and international authorities are continuing to develop testing guidelines. These aim to provide a basis for the production of quantitative data using standardised procedures. These have been successful, in particular, in raising the quality and consistency of industrial product registration submissions to regulatory authorities.

Ecotoxicology has developed rapidly since the early 1970s when the first testing guidelines were developed and now there is considerable experience in performing the tests and in the production of new guidelines.

Statistics has a major role to play in every guideline since the experimental design, data analysis and interpretation of results are central to the conduct of an experiment. The aim of statistics is to help to produce results that are of sufficient quality in the most cost-effective way.

Recently the OECD Test Guideline National Co-ordinators recognised a growing interest within Member countries and the scientific community in issues relating to the use of statistics in Test Guidelines. Aquatic toxicology was identified as the first area to be reviewed with respect to test design and guidance on data analysis. The OECD Environment Directorate approached Shell Research Ltd to carry out a review.

This report presents the findings of that review. The aims are:

(1) to review existing Guidelines with a view to improving the test design and data analysis in individual Guidelines where necessary.

(2) to harmonise data analysis, as far as possible, across Guidelines.

(3) to make general recommendations about the use of statistics, taking into account recent trends in the scientific literature.

The review is basically in three parts. The first part summarises the statistical guidance given in current and draft Guidelines. The second part looks at the scientific literature and discusses the advantages and disadvantages of proposed approaches. The final part gives general recommendations on how the various issues highlighted might be tackled.

2. Review of the Guidelines

The Guidelines reviewed for this report were those either already adopted or in advanced stages of development.

201 : Alga, Growth Inhibition Test (adopted 7th June 1984).

202 : Daphnia, Acute Immobilisation Test and Reproduction Test.
Part I - The 24h EC50 Acute Immobilisation Test (adopted 4th April 1984).
Part II - The Reproduction Test (draft June 1993).

203 : Fish, Acute Toxicity Test (adopted 17th July 1992).

204 : Fish, Prolonged Toxicity Test: 14-day Study (adopted 4th April 1984).

210 : Fish, Early-life Stage Toxicity Test (adopted 17th July 1992).

*** : Fish, Toxicity Test on Egg and Sac-fry Stages (draft March 1992).

*** : Fish, Juvenile Growth Test - 28 days (draft March 1992).

For each of the above a brief summary is given covering only the test design, data analysis and reporting advice contained in the Guideline.

Specific comments are then made for each Guideline. There were many common points that emerged from comparing the Guidelines and these are collected together in section 2.7.

The Guidelines on the Fish Early-life Stage Toxicity Test and the Egg and Sac-fry Stages Test contain virtually identical statistical guidance and have been combined to avoid duplication.

2.1 Test Guideline 201: Alga, Growth Inhibition Test

Design

At least 5 test concentrations are suggested, the range chosen on the basis of a range-finding experiment. The concentrations should be in a geometric series with the lowest chosen to have no observed effect and the highest chosen to inhibit growth by at least 50% and preferably stopping growth altogether. Three replicates are recommended per concentration with 6 replicates for the control. Cell concentrations should be determined at least at 24, 48 and 72 hours.

Analysis

Mean cell concentrations should be plotted against time to produce growth curves for each concentration of the test chemical. Two options are currently available.

(1) Calculate the area under the growth curve and express as a percentage inhibition relative to control. Plot these percentages on semi-log or semi-probit paper against concentration. Fit a line by eye or ‘when a log-normal distribution can be assumed, a computed regression line may be drawn’. Read the EC50 from the graph using the plotted line. The EC50 can be derived for different periods of time (*i.e.* using only part of the data).

(2) Calculate specific growth rates from the slope of the regression line of log(cell concentration) against time. Either use only the first and last times or all time points. Plot the percentage

reduction relative to control against log(test concentration). Read the EC50 from the graph. The EC50 can be derived for different time periods.

There is a statement that the EC50s derived from (1) and (2) are 'not numerically comparable'.

Reporting

Raw data plus graphs. The EC values and NOEC should be reported for each time-point. The methods of calculation must be given.

Comments

(1) The calculation methods (1) and (2) are equal options *i.e*. no preference is expressed. If they do describe different aspects of the test then both are important. However, the Guideline does not give any help with interpretation. It would seem that they might just be variations of one principle. To avoid confusion and to simplify the data analysis it might be advisable to recommend only one.

(2) References are made to other guidelines but the text generally gives no guidance on the application or interpretation of statistical methods or on the experimental design.

(3) There is an unnecessary requirement that the lowest and highest concentrations must produce specified levels of effect (no effect and very large effects respectively). An experiment could yield a precise estimate of the EC50 yet not satisfy these criteria. Such a restriction could therefore be usefully omitted.

(4) The recommendation of geometric concentration spacing may be somewhat restrictive. If the recommended range-finding experiment is able to locate the region of the EC50 with a measure of certainty, then a different experimental design may be used to advantage. Possibly fewer than five concentrations could be required too.

(5) The recommendation that the controls have double the replication of the test concentrations is only an advantage for analyses aimed at determining the NOEC. For EC50 estimation better use may be made of resources by having an extra concentration (see (4) above though).

(6) There is no requirement for a confidence interval for the EC50. This is an unfortunate omission, although deriving one from the suggested graphical analyses would be very difficult.

(7) The percentage inhibitions may be negative and as such cannot be plotted on semi-log or semi-probit paper. There is no guidance on how to handle this situation. To set them to zero might be one option in this case. There is no mention of outliers or what to do with atypical patterns of concentration-response.

(8) Expressing inhibition relative to control may not be the best way of using the control response. This is because using controls to produce percentage inhibition implicitly assumes the control response is a constant when, in fact, it is liable to experimental variation like all the other responses. It is important to understand how the variability of the percentage inhibition changes with the percentage level *i.e.* the mean-variance relationship.

An alternative, and arguably better, approach might be to incorporate the control response directly into a statistical model for the data.

(9) Traditional probit analysis (Finney, 1971) may be inferred by the suggestion to use probit paper or a ‘computed regression line’. Probit analysis would be inappropriate for percentage data. Probit analysis should be applied only to quantal data *i.e.* proportions. If the percentage inhibitions are considered to be proportions out of 100 then probit analysis may well give a good estimate of the EC50, but the confidence interval would most likely give a misleading indication of precision.

(10) The recommendation to obtain EC50s at 24, 48 and 72 hours needs further explanation. If the time-to-response is important then consideration needs to be given to modifying the experimental design and analysing the data in a different way, for example directly modelling the EC50 as a function of time.

2.2 Test Guideline 202: Daphnia, Acute Immobilisation Test and Reproduction Test:

Part I - The 24h EC50 Acute Immobilisation Test

Design

No recommendation is given on the number of test concentrations. Four groups of five daphnids at each concentration and control are suggested as a minimum. The lowest concentration should produce no observable effect, the highest should produce 100% immobilisation and concentrations should be geometrically spaced. Either measured or nominal concentrations may be used for the data analysis. 10% effect in the controls is acceptable.

Analysis

Plot the percentage immobilisation against concentration on log-probability paper. 'Normal statistical procedures' should be used to estimate the EC50 and 95% confidence interval. References are made to other guidelines and a standard text (Finney, 1978).

Where data are inadequate for the standard methods of calculation of the EC50, the geometric mean of the highest concentration causing no immobility and the lowest concentration producing 100% immobility should be used as an approximation to the EC50... the ratio of the higher concentration to the lower should not exceed 2'.

Reporting

The 24 hour EC50, 'preferably' with a 95% confidence interval, 'determined by a suitable method'. 'If possible', the slope of the concentration-response curve should also be given with its 95% confidence limits.

The highest concentration producing no immobile daphnids and the lowest concentration producing 100% immobility are required. The NOEC and LOEC (lowest observed effect concentration) should be reported.

Comments

(1) Probit analysis may be inferred as the recommended method of analysis although no method is explicitly mentioned.

(2) No guidance is given on how to treat immobilised control daphnids in the analysis. It is quite possible, for example, to accommodate control effects into an extension of probit analysis.

(3) Requiring 0% and 100% effects to be demonstrated does not necessarily assist EC50 estimation. Such criteria do ensure that the NOEC and LOEC can be found. As the main aim of the immobilisation test is to determine the EC50 then consideration could be given to removing this restriction.

(4) Methods exist for dealing with situations where straightforward applications of probit analysis fail, for example when 0% responses obtained at low concentrations with 100% response at all high concentrations (Williams, 1986). It is not necessary to resort to using the geometric mean of the lowest concentration giving 100% effect and the highest concentration giving 0% effect.

Part II - The Reproduction Test

Design

The results of the immobilisation test should be used to guide the choice of concentrations. At least five concentrations should be used, chosen in a geometric series. At least 10 daphnids should be held individually at test concentrations and for controls. Different designs are allowed for flow-through systems.

20% control mortality is acceptable. The mean number of surviving offspring per parent surviving in the controls must exceed 60. The coefficient of variation for the reproduction in the controls should not exceed 25%. If a NOEC is required then the lowest concentration should not give a significant reduction in reproduction compared to control.

Test concentrations should be maintained to within 20% of nominal values. Where this is not possible results should be expressed in terms of measured concentrations.

Analysis

Reproduction is measured in terms of mean number of offspring per parent from the start of the test until either the parent dies or the end of the test is reached. Other effects may be analysed for example, survival, time to first brood, number of broods.

The EC50 or NOEC/LOEC for the reproduction per parent may be estimated. For EC50 estimation it is suggested that a suitable concentration-response curve is fitted, possibly weighted to account for differing variability in reproduction at test concentrations. The fit of the model should be assessed.

For NOEC calculation, analysis of variance (ANOVA) may be used if the assumptions behind the analysis can be shown to hold. Data transformation or non-parametric methods are suggested as alternatives. The statistical power of the analysis should be calculated.

It is recommended that a statistician is involved in the analysis of results.

Reporting

A plot of the reproductive output is required, i.e. mean number of young per parent against concentration.

EC50s must be quoted together with confidence intervals. If NOECs/LOECs are reported then the statistical power of the analysis must be given.

Explanations of statistical methods should be given.

Comments

(1) This Guideline gives a reasonable indication of the sorts of analyses that could be considered and recommends that professional statistical advice is sought. Insufficient details are given for those without access to statisticians.

(2) An annex gives details of calculating the statistical power of ANOVA analyses but some may find it hard to follow.

(3) The Part II draft Guideline is a noticeable improvement on the existing Guideline (adopted 4th April 1984) with regard to the standard and clarity of the statistical advice.

2.3 Test Guideline 203: Fish, Acute Toxicity Test

Design

Use range-finding to determine an appropriate concentration range. At least five concentrations are recommended in a geometric series (spacing factor not exceeding 2.2). Use at least seven fish per concentration and for the control. No mention is made of how these should be housed, for example all in one tank or split between more than one tanks. A maximum loading rate is specified.
Control mortality must not exceed 10% or one fish, if fewer than 10 fish were used. The measured concentrations may be used in place of the nominals if these were hard to maintain. Mortality counts are preferably made at 24, 48, 72 and 96 hours.

Analysis

Plot the cumulative percentage mortality for each time against concentration on log-probability paper. 'Normal statistical procedures' should be used to estimate the LC50s and confidence intervals.

Where the data are 'inadequate for the use of standard methods' of calculating the LC50 the geometric mean should be calculated of the lowest concentration giving 100% mortality and the highest concentration giving zero mortality.

Reporting

The LC50s at each time point with confidence intervals should be given. Graphs of the concentration response curves are also required.

Comments

(1) There is no advice or guidance on the statistical methods or experimental design although literature references are cited.

(2) The design recommendations may be too restrictive in some circumstances. In particular the spacing factor for the geometric series could be limiting.

(3) There is no mention of how to deal with control mortality in the estimation of the LC50s.

(4) If a poorly defined concentration-response curve is found, there is no need to abandon standard methods in order to get a satisfactory analysis. Williams (1986) gives a suitable approach.

(5) The requirement for analyses at four time points is not explained and the interpretation of the results at the different time-points is not discussed. If time-to-response is important other features of the design and analysis need to be considered.

2.4 Test Guideline 204: Fish, Prolonged Toxicity Test: 14-day Study

Design

At least 10 fish at each test concentration and for the control. Concentrations must 'permit determination of threshold levels.....and.... NOEC'. The results should be based on the measured concentrations if there is a 20% or greater deviation from nominal levels. Lethal and chronic effects are to be recorded at a minimum of three times per week. No direct mention is made of how to house individuals, i.e. singly or in groups, only a maximum loading rate is given.

Analysis

No details of the analysis are given or references directly cited.

Reporting

Threshold levels of lethal and other effects, together with NOECs should be reported. Results may be presented in graphical form.

Comments

(1) This Guideline offers no guidance on the statistical analysis of the data. Threshold levels are not defined in statistical terms, the decision on whether an 'effect' has occurred is left to the experimenter. It is interesting to note that the definition of the NOEC does refer to statistical significance.

(2) The requirement for the design to be able to determine threshold levels may be interpreted as an acceptance criterion.

(3) No mention is made of control mortality in the analysis or what to do with atypical values.

2.5 Test Guideline 210: Fish, Early-life Stage Toxicity Test and Draft Guideline: Fish, Toxicity Test on Egg and Sac-fry Stages

Design

Five concentrations are suggested, chosen in a geometric series (spacing factor not exceeding 3.2). The acute concentration-response curve should be considered when selecting the concentration range. Narrow concentration ranges may be 'appropriate in some circumstances'. Justification is required if fewer than five concentrations are used.

The number of fertilised eggs at the start of the early-life stage test should be 'sufficient to meet statistical requirements'. No further explanation is given. At least 30 eggs in each of two replicates should be used at each concentration.

There is a recommendation that a statistician is involved in both the design and analysis of the test since the Guideline 'allows for considerable variation' in design and parameters measured.

A randomised block experiment is said to be preferable to a completely randomised design. The random allocation of individuals to treatments is mentioned.

Analysis

The Guideline specifically states that no guidance will be given on test design and analysis because of the number of options available.

Many end-points are suggested each of which might require a particular form of analysis since they encompass a wide range of data types, for example, time, counts, lengths, proportions.

There is suggestion that Dunnett's method (Dunnett, 1955, 1964) might be useful for comparing test concentrations to control.

Reporting

The statistical methods must be reported. The NOEC and LOEC must be given. 'Any concentration-response data and curves available' may also be presented although there is no requirement to do anything with them.

Comments

(1) The advice on the concentration spacing may be restrictive, especially since it is recommended that a statistician is involved anyway.

(2) The stated aim of this Guideline is to enable LOECs and NOECs to be determined. No consideration is given to EC50 estimation, although acute end-points may be observed and concentration-response curves may be reported.

(3) While recommending that a statistician be involved is sensible advice, it seems unfair to omit all details of possible statistical methods. There is no help for those who do not have access to statisticians.

(4) A randomised block design is only preferable to a completely randomised design when there are systematic effects in the laboratory that can be controlled using blocking. Blocking, if used, should be taken account of in the subsequent analysis.

2.6 Draft Guideline: Fish, Juvenile Growth Test - 28 Days

Design

It is recommended that the test design be chosen with regard to the objectives of the study, in particular whether or not the aim is to estimate a NOEC or an EC20 (or other percentile of the concentration-response curve). These correspond to a 'analysis of variance (ANOVA)' design and a 'regression' design respectively. The statement is made that a regression approach is preferred.

For an ANOVA design, a table is given of test designs of different size together with their estimated ability to detect effects of a given magnitude. These were calculated on the basis of earlier ring test results. There is also a discussion of pseudo-replication applied to tanks and fish within-tanks. At least five concentrations are recommended, in a geometric series with a spacing factor not exceeding 3.2. The highest concentration should not be less than 10% or greater than 32% of the 96-hour LC50. It is recommended that twice the number of control replicates are used than for a single test concentration.

For a regression design at least one control tank and five test concentrations are suggested. There is a recommendation that the concentration range should span the likely region of the EC20, a range-finding study possibly being useful in determining this range. Sixteen fish per tank is recommended with one tank per concentration, further tanks may either be used to increase replication or increase the number of concentrations.

Three time periods are identified, 0-14 days, 15-28 days and 0-28 days. Fish weights are to be determined on days 0, 14 and 28.

Analysis

The analysis is based on the specific growth rate. This may be defined in various ways depending on how the measurements on fish are taken. The analyses are to be carried out for each time period.

ANOVA, leading to NOEC/LOEC estimation is discussed with a suggestion that Dunnett's or Williams' (1971, 1972) methods may be used to make comparisons of test concentrations with control.

For regression analysis, a simple linear model is suggested unless the data indicate non-linearity. In this case non-linear least squares may be used to fit an appropriate model. A confidence interval for the

EC20 must be calculated. If mortality is present then a weighted regression is suggested to take account of different numbers of fish in each tank.

Reporting

Either an EC20 or NOEC may be reported for each period.

Comments

(1) There is good discussion of the two approaches to the design and analysis compared to the other Guidelines. Pseudo-replication is discussed. The preference for regression is clearly stated, although either analysis is acceptable for reporting.

(2) No mention is made of outliers or how to handle to control mortality, or mortality at test concentrations (other than the comment about a weighted regression analysis made above).

(3) The prescription of a geometric spacing factor for the concentrations not to exceed 3.2 could be too restrictive.

2.7 Common issues and summary of Guideline review

There is a balance to be struck between making a Guideline an almost rigid protocol and giving experimenters and statisticians the freedom and flexibility to use their knowledge and experience in the design and analysis of any particular study.

While it is accepted that Guidelines are only for guidance, they tend be followed very closely. The comments below are made in the spirit that people will abide by the Guideline recommendations wherever possible.

Complete specification of the statistical design and analysis, sufficient for use in all situations, would not be possible. Furthermore it would not be optimal since any prior knowledge the experimenter may have can be used to tailor the experimental design to each situation. Similarly, no one statistical analysis will be correct all of the time.

For the statistical elements of the Guidelines it is therefore argued that the philosophy should generally be to try and ensure maximum freedom and flexibility. Recommendations on options for the design and analysis may come from consideration and analysis of available data (e.g. from ring tests). This would then ensure that experimenters without access to statistical advice will always have sufficient guidance. Statisticians will be made aware of the limits within which it is possible to report alternative analyses.

It is felt that the present Guidelines generally offer too little flexibility in the choice of experimental design and too little help with the statistical analysis.

Below, the general issues arising from the Guidelines are discussed, divided between design, analysis and reporting. They are not in any order of importance. Some of the issues highlighted will be covered in more detail in sections 3.3 and 3.4.

Some of the issues raised below are somewhat open-ended since they give rise to questions that need to be addressed by aquatic toxicologists in connection with specific Guidelines.

These discussions inevitably imply recommendations for change. How best to bring about change will be outlined in section 4.

Design

(1) The recommendations for the experimental designs could be too restrictive.

A typical requirement for five concentrations in a geometric series with a limit on the spacing factor could be relaxed. For example, given good prior knowledge of the possible location of an EC50, an experiment with just three concentrations might be quite adequate to obtain a precise estimate. An experimenter may have such information from previous experience with similar species of organism or similar chemicals.

A geometric series, while generally very good, may not be optimal for the estimation of an EC50 in all circumstances. Bounds on the factor spacing, for example 2.2 or 3.2, may also be too limiting.

It is recognised that some suggestions for the design need to be made to assist those without access to professional statistical advice. However, a range of flexible options should be offered. Tables of designs could be produced, as for the Fish Juvenile Growth Test.

(2) The statistical power of the analyses should be considered more carefully.

The numbers of replicates and numbers of individuals at each replicate should be chosen to provide a required level of sensitivity for the statistical comparisons of test concentrations with control when NOEC estimation is required.

Few of the Guidelines discussed sensitivity i.e. what size of effect should an experiment be able to detect and with what degree of certainty. There were also no general requirements for experimenters to report the sensitivity of their studies.

This issue is very important, since more variable data implies lower sensitivity and will generally lead to higher NOECs. This, in turn, implies greater apparent environmental acceptability.

(3) For ANOVA, it is usually preferable to have greater replication for controls than for test concentrations.

For the typically-sized experiment being considered here, double the number of control replicates is advisable. The small increase in work is repaid by much greater sensitivity for the comparisons of test concentrations with control. This was over-looked by several of the Guidelines.

(4) The Guidelines do not give sufficient attention to the issue of pseudo- replication and multiple housing of individuals.

Where individuals are kept together in a tank or vessel then, strictly speaking, the tank is the experimental unit, not the individual organism.

How this is handled in the statistical analysis can have a marked effect on the apparent sensitivity of the statistical tests or the precision of an EC50. The problem is more acute for NOEC estimation since changing the sensitivity of the statistical analysis may result in different values of the NOEC. An EC50 estimate will, however, be largely unaffected.

In some circumstances it can be shown that the tank-to-tank variability is small compared to the variability between individuals. If so then it can be argued that the individuals can be used as replicates. However many statisticians might disagree with this. Such analyses could be usefully carried out on ring-test data, as with the Fish Juvenile Growth Test, to provide a 'definitive' answer for each Guideline where the issue might arise.

Some Guidelines mentioned limits on tank loading rates for fish, for example. There were no comments on how to split the numbers of individuals if the suggested loading rates were exceeded. This would clearly have implications for the statistical analyses and warrants further consideration.

(5) The use of measured or nominal concentrations in subsequent analyses needs to be clarified.

If nominal concentrations are used to determine, for example, an EC50 yet measured concentrations deviated from nominal levels then a bias could be introduced into the EC50 estimate. Such a bias may or may not be biologically significant. The bias may be very important if the EC50 is close to a toxic classification boundary value. The implications should therefore be considered separately for each Guideline.

The preferred situation might be to always use measured concentrations. It is recognised that this may have significant cost implications.

Another issue arises with replicate tanks. If the measured concentrations in replicates differed markedly then, for an ANOVA analysis, they would cease to be genuine replicates. This would present no problem for a regression analysis however since different tanks would simply be regarded as different test concentrations.

(6) Acceptance criteria are not required to validate the statistical analyses.

Most of the Guidelines include acceptance criteria. Often these are based on control effects. It should be noted that such criteria are generally not necessary to validate the statistical analysis. The sensitivity of analyses may be affected but probably not markedly.

(7) Requiring 0% and 100% effects to be demonstrated is not necessarily of benefit to the statistical analyses.

For EC50 estimation there is no good reason to request that such effects are obtained. It is more important to ensure that the main part of the concentration-response is adequately characterised. It is possible that experimental effort could be saved if this requirement is relaxed.

For NOEC estimation, requiring a 0% effect will ensure that a NOEC can be obtained. Therefore this requirement is almost acting as an acceptance criterion.

(8) The need to randomise the experiment was not adequately stressed.

Randomisation affords protection against unknown systematic biases and should be employed whenever possible. Practical constraints sometimes outweigh the statistical arguments but caution should be exercised and consideration given to randomising all elements of an experiment.

(9) Systematic effects in the laboratory should be investigated.

The need to look for any systematic effects in the laboratory was not explicitly stated. Local environmental effects do exist, such as light and temperature variation. These can be controlled by blocking (Cox, 1958).

Blocking was mentioned but it was stated that blocking is preferable to a completely randomised design. This is true only if systematic effects do actually exist. Otherwise unnecessary degrees of freedom are used up in subsequent analyses. The need to account for blocks in the analysis was not stated.

Analysis

(1) A graphical summary of the concentration-response curve should always be requested.

This is the simplest way of showing the relationship established by the experiment and, by including replicates, it is immediately obvious what the degree of scatter is in the data. Outliers are also readily identified.

(2) The end-points that can be analysed need to be clearly stated and defined.

Some Guidelines suggest that a list of end-points might be analysed, others that such end-points might be analysed if the experimenter thinks them important. The number of these end-points should be minimised. The temptation to analyse everything that could be analysed should be strongly resisted. Simplicity should be maintained as far as possible.

Also, different types of end-points (e.g. proportions, time etc.) may need different types of statistical analysis.

(3) Graphical methods for EC50 estimation should be discouraged.

For example, some Guidelines prescribe graphical methods of determining EC50s, including drawing probit lines by eye on appropriate graph paper. Modern technology largely removes the need for this as a first choice method.

It is quite likely that estimates of EC50s obtained by eye will be close to those obtained using probit analysis or non-parametric methods. This is because the EC50 is a very robust summary statistic. However, the main criticism is that confidence intervals are hard, if not impossible, to calculate from a graphical analysis.

Non-parametric methods of estimating EC50s and confidence intervals are often very simple to operate on a hand calculator even if there is no access to more advanced computing. This is the very least that should be expected.

(4) It is essential that EC50s, or any other percentiles, are accompanied by confidence intervals.

If confidence intervals are not given then there is no way of knowing how reliable the estimate is. The width of a confidence interval reflects the experimental design, conduct of the experiment and the biological variability.

(5) Probit analysis can still be used when the concentration-response curve is poorly defined.

Many standard implementations of probit analysis will fail if the responses at low concentrations are all 0% and are 100% at higher concentrations. Similarly, if there is just one concentration giving a response between 0% and 100% then this may lead to numerical problems.

Williams (1986) and Van der Hoeven (1991) describe methodology that can be used to derive valid confidence intervals in these situations. Therefore there is no need to resort to calculating the geometric mean of the lowest concentration giving 100% and the highest concentration giving 0% response. Such an approach does not enable a confidence to be calculated either.

(6) Guidelines offered no advice on how to use control mortality data in the analysis of acute studies.

How to handle control effects was generally ignored. Options are to ignore it, adjust the data or to incorporate control effects directly into a model for the data. Adjusting the data is somewhat ad hoc, is a legacy of the pre-computer era and should be avoided.

Either of the other options could be acceptable depending on a laboratory's previous experiences. For example, if control mortality is seldom seen then ignoring the odd occurrence in the analysis presents no problem and the control effects can simply be noted as coincidental findings.

If control mortality is seen more regularly then this indicates that non-treatment related effects may also be occurring at test concentrations. In this case it is preferable to make allowance for this in the concentration-response model.

(7) The issue of outliers was not addressed.

Statistical tests exist for detecting outliers but often a plot of the data is often sufficient to enable atypical points to be identified. Also it would not be desirable to omit points solely on the basis of statistical tests. The experimenter should always have the responsibility for justifying the removal of data.

An alternative might be to report analyses both with and without suspect points. As already noted, it is possible that small changes in findings could be important if the values are close to cut-off values for toxic classifications.

A further alternative would be to use outlier-resistant methods. Non-parametric methods are one possibility here, but these may not be sufficiently flexible for some modelling needs, for example, handling control mortality. Robust regression methods seek to down-weight the influence of atypical points in model fitting. The modelling flexibility would be maintained, however there is an extremely wide range of options available.

(8) Analysing results expressed as a percentage relative to control is not recommended.

While this may be convenient it may lead to problems. For example, negative values may be obtained. No comment was made on this in any Guideline. For statistical modelling it is better not to make this transformation but to model the raw data and incorporate the control response directly.

(9) Probit analysis is not the only concentration-response modelling technique available.

Probit analysis is only strictly appropriate for quantal data i.e. proportions. Use of probit analysis for the Algal Growth Test may lead to misleading estimates of the confidence interval. The EC50 would most likely be well estimated.

There are several issues to be examined when using concentration-response models. The most important are the form of the model, the model parameterisation, the mean-variance relationship and the method of model fitting. These are technical details and really require statistical advice, at least initially.

There were generally no requirements to demonstrate that a probit, or any other model, actually fitted the data satisfactorily. Formal statistical tests should be employed to check model fit or graphs of suitable diagnostics (e.g. residuals) could be used.

It is unreasonable to expect experimenters to be able to address all of these technical matters in general. It is preferable that sufficiently detailed recommendations are made in the Guideline. Investigation of these issues and subsequent recommendations could still leave a good degree of flexibility should alternative analyses be warranted.

(10) The advice given for carrying out ANOVA did not cover the technical details to an acceptable depth.

To estimate NOECs and LOECs, multiple comparison procedures are used to compare each test concentration to control. There are very many of these. Only the well-known Dunnett's and Williams' methods were referred to.

There were generally no requirements to show that the assumptions behind the ANOVA were met, for example, normality and homogeneity. This is an important omission. There was also no advice given on the possible need to transform data to satisfy the assumptions, yet these methods are described in most basic statistics textbooks.

The use of non-parametric methods was not discussed. These might provide, arguably, less sensitive analyses but will be robust against failure of the usual assumptions.

(11) It is debatable whether an extensive list of statistical references should be given in each Guideline.

It might be just as useful to present a worked example that brings out all the relevant points that should be considered.

(12) Time-to-response could be incorporated directly into the analyses.

Several Guidelines mention performing analyses at different time-points. If time-to-response is an important feature in assessing toxic effect then consideration could be given to specifically incorporating time as a factor in the analyses. Both ANOVA and regression modelling could be modified.

Statistically, the issue is that the data at different time-points are correlated since the same individuals are being observed. Making use of this will not only make the analyses more statistically correct but may also improve precision.

With regression modelling there is the further possibility of modelling the EC50, say, as a function of time. This would enable the EC50 to be predicted at time-points other than those tested, together with confidence intervals. If such predictions were of value then, ideally, the experimental design of these studies should be re-examined.

Reporting

These comments bring together some of the points already made above.

(1) A graph of all the data should always be reported.

This is the best way of showing the concentration-response. All data points should be plotted, not just replicate means.

(2) Confidence intervals should be reported for all summary statistics.

For example an EC50. Note that a confidence interval cannot be reported for a NOEC. However, the power of the tests could be quoted and the magnitude of effect that could be detected.

(3) A statistical interpretation of the results may be helpful.

This could draw attention to any special features of the data and the analysis, for example outliers.

3. Statistical methods

This section reviews the literature on statistical methodology pertaining to aquatic toxicology and discusses some of the points that emerge.

The two main options available for the analysis are ANOVA and EC point estimation. These are discussed in sections 3.2 and 3.3 respectively.

The experimental design is closely linked to the method of analysis and features relevant for both ANOVA and EC point estimation are covered in section 3.4.

Finally, in section 3.5, conclusions are drawn as to the preferred method of analysis.

3.1 A review of the scientific literature

It was mentioned above that there are basically two main types of analysis that have been proposed.

The first is ANOVA, principally leading to the estimation of NOECs and LOECs. The NOEC is an extremely well established summary of chronic toxic effects and is recommended in other international guidelines, for example EPA. It is determined by comparing the responses at each test concentration to those of the controls using some statistical procedure, typically known as a multiple comparisons procedure.

Definitions may vary slightly, but the NOEC is usually taken as the highest concentration that is not statistically significantly different from control. The LOEC is similarly defined as the lowest concentration that shows a significant difference from control. A further statistic may be defined. This is the maximum acceptable toxicant concentration (MATC) and is the geometric mean of the NOEC and LOEC.

The second method of analysis is to estimate an EC*x* (where *x* is the percentage point on the concentration-response curve, e.g. 50). This is routinely done for acute toxicity where the EC50 is almost universally required by Guidelines. An EC*x* is often required for chronic data but a NOEC is also typically requested too. Probit analysis is just one statistical method for estimating an EC*x* for quantal data. Non-parametric methods are very popular as they avoid some of the problems of concentration-response modelling in certain situations.

Closely related to an EC*x* estimate is the so-called 'benchmark' concentration. Crump (1984) effectively defines this to be the lower confidence limit for the EC*x*. Thus the benchmark concentration may be seen as a conservative estimate of the EC*x*. Hoekstra and Van Ewijk (1993) propose a similar statistic they call the 'bounded-effect' concentration.

It could be argued that there is some inconsistency between the requirements for statistical analysis of acute and chronic tests. This is because the ANOVA approach need not acknowledge the existence of a concentration-response relationship. EC*x* estimation, however, may make explicit use of a model of the toxic response curve. In the interests of simplicity and consistency it seems reasonable that a common framework for the analysis of both acute and chronic effects should be sought.

The reason for the distinction between acute and chronic statistical analysis probably lies in the notion of a 'safe' level. ANOVA naturally leads to pairwise comparisons with control and the testing of each concentration for its 'effect'. With EC*x* estimation there is an acknowledgement that any concentration, however low, will produce some effect although this may not be observable.

There is another argument. This is that there is a toxic 'threshold' i.e. there is a maximum safe level below which there are no adverse effects whatsoever. Above this threshold some concentration-response relationship will hold.

This argument is often put forward as justification for the ANOVA and NOEC approach and used to reason that EC*x* estimation is not appropriate. This is not a valid defence of ANOVA but is a valid criticism of EC*x* estimation. It is, in fact, possible to make use of threshold models to derive a statistically valid estimate of a genuine no-effect-concentration (NEC). This approach is not without statistical problems either but is certainly preferable to ANOVA if a threshold is to be assumed and the principle can be applied to chronic and acute data if necessary.

It is also useful to consider the role that trend analysis might have. Williams' method for comparing concentrations with control makes the assumption of a monotonic response relationship and is often used to determine the NOEC.

Real evidence for the existence of toxic thresholds might be difficult to find and would probably be hard to prove statistically. This because very large experiments would be needed to statistically discriminate between the different models.

It is important to note that none of the issues being discussed actually involve technical mis-use of statistical methods. ANOVA and regression may be applied perfectly correctly. What is important is the interpretation placed on the results.

Skalski (1981) argued that there is logical and statistical flaw in the way the NOEC is interpreted since one can never prove that the null hypothesis being tested, i.e. that there is no effect, is actually true. The emphasis should be on disproving the presence of effects, a point also made by Hoekstra and Van Ewijk (1993). Skalski (1981) also concluded that the NOEC did not provide the necessary information from which to determine a safe concentration.

Stephan and Rogers (1985) drew attention to the statistical weakness in using the NOEC as a 'safe' level. The basic argument is that the NOEC is totally dependent on the design and conduct of the experiment. They strongly suggest that concentration-response modelling should be used as this avoids most of the key problems of ANOVA and gives 'unbiased' estimates of toxic end-points. The choice of effect level is a question for toxicologists.

Others have discussed statistical drawbacks to the NOEC. Masters *et al* (1991) noted the experimental dependence of the NOEC and suggest that multiple comparisons procedures have been both over-used and mis-used. They favour the use of biologically-based concentration-response models. Bruce and Versteeg (1992) also pointed out that information about the concentration-response is lost when the data are summarised by the NOEC.

Leisenring and Ryan (1992) investigated the statistical properties of the NOEC under the assumption of an underlying Weibull concentration-response model. Their findings confirmed the sensitivity of the NOEC to the size of the experiment and the variability. Furthermore, they found that the NOEC could correspond to substantial effects. They concluded that the shortcomings of the NOEC provide an incentive to develop model-based methods for expressing the relationship between concentration and effect.

Barnthouse *et al* (1987) and Suter *et al* (1987) both commented on the fact that 'large' effects may be present at 'safe' concentrations. This is a direct consequence of the statistical power of the testing procedures. This, in turn, is a consequence of the design of the study and the variability seen in the responses. Liber *et al* (1992) found they could only detect effects of nearly 50% in some of their mesocosm studies. Oris and Bailer (1993) drew the obvious conclusion that any experimental design must take into account the ability to detect effects of specific sizes.

McClave *et al* (1981) and Feder and Collins (1982) discussed the design of experiments and express dissatisfaction with ANOVA analyses. Crossland and LaPoint (1992) noted another feature of design. With a regression design they commented that a greater concentration range can be studied than with an ANOVA design of the same overall size.

Criticism of the NOEC is found in areas other than aquatic toxicology. Crump (1984) and Gaylor (1989) both discussed the NOEC and its uses in mammalian toxicology and risk assessment. Crump noted that the NOEC is not an inherent property of the test species but is heavily dependent on the experimental design. Gaylor, too, noted the dependence on the experimental design and pointed out that the NOEC cannot represent a safe level as the experiment may have been too insensitive to detect subtle effects.

In the agriculture literature, Cousens and Marshall (1987) recommended that multiple comparisons should not be used in experiments where there is a relationship between treatments i.e. a range of concentrations of one single chemical. A clear statement about the acceptability of multiple comparisons comes from Buxton (1982) who, in editorial instructions to authors of an agronomy journal, stated that they should not be used where the treatments are related in this way.

Alternative analyses have focused on either the estimation of EC*x* points or benchmark concentrations. A key issue is what EC point to estimate. Bruce and Versteeg (1992) say that an EC20 is common. Barnthouse *et al* stated that ecological risk assessment does not involve extrapolation to extremely low concentrations and offer the EC25 as a reasonable statistic. Nyholm *et al* (1992) in turn based much of their regression analysis around an EC10 and seem to suggest that this might be an approximation to a toxic threshold. Finally, Westlake *et al* (1983) suggested that a 5% effect might be considered 'safe' for Daphnia reproduction. The issue of what size of experiment is needed to estimate such effects with acceptable levels of precision has received little or no attention.

Oris and Bailer (1993) commented that an LC25 is often close to the NOEC but they rightly note that this is likely to be due to the lack of sensitivity of the statistical analyses rather than anything more fundamental.

Some authors have tried to compare EC50s with NOECs. Oris *et al* (1991) concluded that both EC50s and NOECs are needed. Sloof *et al* regressed log(NOEC) on log(EC50) and then attempted to extrapolate some of their findings across species and to the ecosystem. DeGraeve *et al* (1992) compared EC50s from a ring test between and within laboratories but still use NOECs to quantify certain aspects of reproducibility. EC*x* values and NOECs, as commonly estimated, can never be compared and nothing of use can come from trying to establish relationships between them.

Mayer *et al* (1986) used elaborate statistical methods to compare NOECs between different studies involving different chemicals and different fish species. Again, the design dependence of the NOEC means that such an investigation cannot be relied upon to provide sensible conclusions.

Hoekstra and Van Ewijk (1993) discussed a way of estimating a conservative benchmark concentration using methodology similar to Crump (1984). They point out that 'classical' concentration-response modelling is unreliable when low percentiles, for example EC10 or smaller, are to be estimated. Hamilton *et al* (1977) also made this point and generally favour non-parametric methods. However, they acknowledge that non-parametric methods have drawbacks when trying to estimate small effects and lack flexibility to handle control mortality and hormesis (*i.e.* slight stimulation of response at low concentrations).

The calculation of confidence intervals for EC*x* values has not received the attention that it merits. Walsh *et al* (1987) made the surprising statement that the 'calculation of confidence limits from quantitative data is questionable'. By their nature benchmark estimates do not have associated confidence intervals and this must be a drawback.

Van der Hoeven (1991) promotes a sound method of calculating confidence intervals using maximum likelihood. As this is computationally involved an extensive table is provided that deals with a special type of problematic concentration-response curve. Williams (1986) provides further details set in a more general dose-response setting.

Bruce and Versteeg (1992) sensibly reparameterised their model to make the EC*x* a parameter in its own right. This has the advantage that standard computer programs for fitting non-linear models will automatically provide estimates of error for the EC*x*. This is preferable to using approximations to the variance of some function of the parameters for example the well known Fieller's theorem. Nyholm *et al* (1992) did not use this simple manipulation for their models.

Threshold models have not been studied to any great extent in the aquatic toxicology area. Hoekstra and Van Ewijk (1993) noted that estimating the threshold value will typically be a very poorly conditioned problem and, for the typical experiment, will lead to very imprecise estimates that are highly model dependent. Cox (1987) looks at threshold models in toxicology.

Other authors have used a variation on regression methods to estimate a threshold concentration (NEC). Liber *et al* (1992) used a straight line model and looked to see where the confidence interval about this line intersected with the confidence intervals for the control mean. As straight line models are rarely appropriate for concentration-response curves, it is unclear how sensitive their approach might be to choice of model. They also present a method of determining a confidence interval for the NEC although it is not clear how successfully this will operate compared to alternative methods. Capizzi *et al* (1985) presented a complex sequential testing procedure assuming an underlying model.

From the literature it is clear that the NOEC concept is open to criticism. Discussion of the drawbacks of NOEC have become more common in recent years but objections were raised much longer ago. The basic issues have also been covered in other areas of science. The fact that the NOEC is still widely used must, in part, be due to the difficulty in developing and adopting alternatives.

However, there is an increasing amount of work being published on modelling and other approaches. The alternatives on offer either aim to estimate an EC*x* or estimate an equivalent benchmark concentration. Much less work has been done on estimating NECs using threshold models.

3.2 The ANOVA/NOEC approach

Some of the literature cited above make many of the points below. In particular, Stephan and Rogers (1985) presented many of the problems. However, their discussions may not be readily accessible to non-statisticians.

Potential problems with the NOEC are demonstrated in a series of points below.

(1) The NOEC must be one of the test concentrations.

This is obvious but the implication is that the NOEC is determined by the experimenter's choice of concentration levels. As concentrations are often geometrically spaced and there may typically only be five of them, then it is hard to see how the NOEC could generally be considered as an accurate estimate of a 'safe' concentration.

(2) No precision statements are possible for the NOEC.

If two laboratories were to report different NOECs for the same chemical there is no way of knowing which is the more reliable. At the very least, experimenters should be encouraged to report the basic error variability of their data or what size of effect they could detect, however this is seldom done. Good statistical practice requires that summary measures should come with an associated confidence interval (error bar). None can be calculated for the NOEC by definition.

(3) NOECs may correspond to large effects on test organisms.

If the variability in an experiment is relatively high then the corresponding sensitivity of the statistical analysis will be relatively low. Only large differences from control could then be detected. Consequently, the resulting NOECs could themselves correspond to large and potentially biologically important magnitudes of effect.

(4) The NOEC will not be obtainable in all cases.

If the lowest concentration produces a statistically significant effect then the NOEC will not exist. Similarly, if there is a very slight trend across the concentrations then the NOEC could be the highest concentration and experimenters may be concerned that they have not found the 'toxic threshold'. Both of these outcomes may lead to a repeat experiment with inherent cost penalties and the use of more test organisms.

However, it is clear that both experiments have yielded much information about the toxic response and alternative methods of analysis can extract such information. The experiment need not necessarily be repeated.

(5) ANOVA on its own may be wasteful of information.

This is because it fails to provide a quantification of the slope of the concentration-response relationship i.e. the range of sensitivity of the test organism. General trend analysis, linked to ANOVA, will also not fully describe the range of sensitivity.

The above points indicate that the NOEC is far from ideal as a summary measure of toxic effect. It is too heavily dependent on the experimental design and the variability in the data. Consequently the NOEC may correspond to large effects, possibly of biological significance. Its value in hazard assessment is questionable.

3.3 EC point estimation

EC point estimation does not share the problems of the NOEC. However, EC point estimation is not without difficulty. No analysis can be perfect so the question is one of balance and whether the advantages outweigh the disadvantages.

In this section estimation of an EC*x* will be considered together with the idea of benchmark concentrations (e.g. Hoekstra and Van Ewijk, 1993) since all of these methods basically attempt to estimate concentrations producing specified effects. The generic term EC*x* will cover all such estimators.

EC*x* estimates can be made using a wide variety of techniques. It is convenient to divide the methods into model fitting (or regression) and non-parametric methods.

Regression involves fitting a concentration-response curve to the data and interpolating or extrapolating to obtain the EC*x*. A particular subset of regression models are threshold models which contain a discontinuity at the point of toxic threshold, the NEC. Models should ideally be biologically plausible rather than just provide a good fit to the data points. However, 'usefulness' of the model is arguably more important than plausibility.

Non-parametric methods do not postulate a model. Instead, the observed data may be smoothed and combined to estimate an EC50, for example. The well known Spearman-Kärber and moving-average angle methods are examples of smoothing methods. Other methods involve ranking the data (R-estimators) or using linear combinations of order statistics (L-estimators). These latter two techniques do not seem to have found much use in aquatic toxicology to-date yet may offer some advantages in terms of greater robustness and the fact that no choice of model is required. Morgan (1992) provides an excellent review of methods for quantal data (proportions) which is therefore relevant for acute studies. Generalisations to other toxic end-points are possible.

Advantages

(1) The ECx is not restricted to be one of the test concentrations.

The EC*x* may take on any value. Because the EC*x* is interpolated it is not 'biased' by the choice of test concentrations and will be relatively insensitive to the number and spacing of concentrations, compared to the NOEC.

(2) The precision of the ECx can be quantified.

Thus less noisy experiments will generally be rewarded by greater precision and therefore narrower confidence intervals.

(3) ECx values are comparable.

Taken together with associated confidence intervals, EC*x*s can be meaningfully compared between repeat experiments or between laboratories, where the same test protocol has been followed. This is really a consequence of points (1) and (2).

The point estimate of the EC*x* will be less sensitive to the variability in the experiment than the NOEC. Any variation between experiments will be shown by the width of the confidence intervals. It is then necessary to compare EC*x* estimates relative to their precision.

(4) Measured or nominal concentrations can be used without any statistical problem.

With EC point estimation, the measured or nominal concentrations can be used whether or not these are different between replicates. For ANOVA differences mean that replicates are not true replicates in the statistical sense. For regression, measured concentrations can be treated simply as different concentrations in their own right.

(5) The whole of the toxic response of the organism may be characterised.

This will be the case when a model is fitted to the data. Non-parametric methods do not allow this in general.

In particular, the slope of the concentration-response curve conveys much information on the sensitivity of the organism to changes in concentration. Also, effects can be estimated for concentrations other than those tested and confidence limits determined.

(6) Regression modelling is flexible.

Model fitting offers a way of handling non-standard concentration-response problems, for example, hormesis. In acute studies, control mortality can be handled easily by minor modifications to the model.

(7) Pseudo-replication is not a crucial issue.

The degree of replication does not play such a vital part in the point estimate of the EC*x* as it does for the NOEC.

Replication is still highly desirable as this enables the intrinsic variability of the experimental material to be quantified. This is important to ensure that confidence intervals accurately reflect the degree of uncertainty.

(8) A greater concentration range can be studied.

The degree of replication is less important for EC*x* estimation than for NOEC estimation, where sufficient degrees of freedom are needed to achieve acceptable levels of statistical power.

Therefore for the same overall size of experiment, experimental material can be redeployed to explore a wider concentration range if desired.

It is possible that EC*x* estimation would result in less costly or more cost-effective experiments as the effort could be reduced in many cases. See section 3.4.

(9) Acute and chronic studies can be analysed using the same basic approach.

This is because some form concentration-response analysis would be performed for each.

Disadvantages

(1) The difficulty in choosing a model.

This is usually raised as an objection to regression modelling. However, it is often overlooked that there are probably more multiple comparison methods available for deriving the NOEC than there are models that could be used for regression. Similarly, proponents of non-parametric methods argue their case from the point of view of robustness. However, there are very many non-parametric techniques too.

Therefore model choice is no more a problem than choice of method for any other statistical analysis.

There are more important issues to do with model choice that need to be discussed. Firstly, the validity of a model can be checked. This can be done using statistical tests or by calculating suitable diagnostic statistics based on residuals for example. There are several options available and some or all should always be used to justify a chosen model. Cook and Weisberg (1982) give a good review of methods for linear regression and these can be modified for the non-linear case.

Secondly, although point estimates of an EC*x* will vary between models, the confidence intervals about these estimates will usually overlap to a considerable degree. Therefore the precision of the EC*x* can be seen to be relatively insensitive to the model. This is one reason why it is so important to always quote confidence intervals with an EC*x*.

Model sensitivity is most noticeable when very low (for example, EC10 or below) or very high percentiles are estimated. Different models may produce quite different EC*x*s. However, in these cases the confidence intervals will become increasingly wide and the overlap will still be great.

In conclusion, model choice need not be a major argument against regression analysis.

(2) *For extreme percentiles confidence intervals may be very wide.*

Extreme percentiles may be taken to be EC10 and smaller or EC90 and greater.

This leads people to dismiss such analyses as providing little 'hard' information. However, wide confidence intervals are simply a reflection of the real lack of knowledge about the particular region of the concentration-response curve. This seems to be simply a fair reflection of the data.

If low percentiles are required then experiments need to be approached in a different way, probably involving greater resources.

(3) *ECx estimation is generally computationally more difficult than NOEC estimation.*

Exceptions to this are many non-parametric methods that can be obtained using a hand-calculator.

Regression modelling, particularly non-linear regression, is more difficult than ANOVA. However, non-linear model fitting routines are contained in many standard computer packages so access should not be too problematic. Use of non-linear modelling does require more care and this is certainly a major drawback of such methods.

It is argued that ability to carry out a statistical analysis should not prevent its use if it is thought to be the best option. The cost of carrying out even a difficult statistical analysis is trivial compared to the cost of performing the experiments in the first place. Computational methods may, of course, be automated and tailored to specific needs.

(4) *ECx estimates may be difficult to obtain in some cases.*

A typical example of this is when low concentrations give 0% response and high concentrations give 100% response with no intermediate responses at any concentration. Also, the case of a single concentration giving an intermediate response between 0% and 100% can cause computational problems.

This is used as argument in favour of non-parametric methods that do not explicitly fit a model. However, it should be recognised that although the point estimate of the EC50 may be indeterminate for the above cases, a valid confidence interval can still be obtained. Williams (1986) and Van der Hoeven (1991) use maximum likelihood to derive likelihood ratio confidence intervals. This methodology can be modified for any non-linear model. Crump (1984) argues along similar lines.

(5) Using ECxs in place of NOECs requires the value of x to be specified.

Thus the emphasis is put on ecotoxicologists to quantify biologically important levels of effect. This may be difficult to do.

Stephan and Rogers (1985) have no hesitation in insisting that ecotoxicologists must think in terms of effect levels. From the literature it would appear that many ecotoxicologists consider effects in the range 5-20% to be biologically acceptable depending on the species involved and the type of effect.

The choice of percentile is not a statistical issue but does impact on the use of statistical methods, principally regression modelling.

(6) Use and understanding of precision and confidence intervals must be increased.

The point estimate of an EC*x* is not sufficient on its own. Regulators and those involved in risk assessment need to make greater use of the quantification of uncertainty associated with EC*x*s.

Other issues

(1) Threshold models avoid the need to specify what ECx value is required.

These models explicitly contain a NEC parameter as the toxic threshold. However, these models may contain more parameters than a 'standard' concentration-response model and may be difficult to fit to typically-sized aquatic toxicology experiments.

Cox (1987) discusses and compares a variety of models and cautions that care should be taken in interpreting the estimated threshold. He also commented that larger (i.e. more concentrations) experiments are really required to be able to use threshold models satisfactorily. This is so that the region where the threshold lies is well located. For aquatic toxicology experiments this would probably mean having more very low concentrations than at present. These may be difficult to maintain.

(2) There is a philosophical difference between threshold models and standard concentration-response models.

Threshold models assume that there are no toxic effects below a certain concentration. Probit models, for example, assume that there are effects, however small and undetectable these might be, at all concentrations.

For small experiments it will generally be very difficult to distinguish between the quality of the fits of these models on purely statistical grounds so the choice may well rest with ecotoxicologists and their degree of belief in one model or another.

In view of the desire to estimate 'safe' concentrations, more theoretical and practical statistical analysis should be carried out to gain an understanding of threshold models in ecotoxicology.

(3) Non-parametric methods may lack flexibility.

They are readily available for estimating EC50s in standard situations. However, little seems to be available for dealing with hormesis or control mortality in acute studies. Ad hoc corrections may be possible for data with control mortality but may not be fully satisfactory.

A further problem will be in estimating low percentiles. Little work appears to have been done on how non-parametric methods perform at the extremes of the concentration-response curve.

(4) Non-parametric methods may be wasteful of information.

One example of this will be in cases where a parametric regression model provides a good description of the data. A consequence of this is that better precision is usually obtained when a good model for the data can be found.

3.4 Experimental design

In aquatic toxicology, experimental design is concerned with the number of concentrations, the concentration levels and the number of replicates at each concentration. The design should be tailored to estimate the statistic of interest with the greatest possible accuracy and precision.

This aim needs to tempered by the practical constraints of experimentation. These will include the ability to maintain exposure concentrations and the physical constraints on the number of concentrations and replicates that can be handled. Cost is also a very important factor to consider.

Statistical analysis can be used to assist the experimenter with the design in a wider sense than indicated in existing Guidelines. ANOVA and EC point estimation make different requirements on the design for optimum performance although there is nothing to stop a good regression analysis being performed on a design intended for ANOVA. The converse may not necessarily be true.

ANOVA

(1) It is necessary to decide exactly what constitutes a replicate.

It may be permissible to use pseudo-replicates as replicates. For example, consider the case of fish within tanks. If the tank-to-tank variability was very low compared to the fish-to-fish (within tank) variability then it could be argued that the individual fish could be treated as replicates rather than the tanks. However, this would be a somewhat controversial step that many statisticians would not accept. Low, relative, tank-to-tank variability was, indeed, found for the Fish Juvenile Growth Test.

(2) Control replication should exceed that for test concentrations.

For a given, fixed overall size of experiment, more sensitive comparisons between the test concentrations and control are made if the control replication exceeds that for the test concentrations. It can be shown that a factor of √(number of concentrations) greater replication for the controls gives optimal sensitivity.

(3) The chosen concentration levels determine the possible NOEC values.

Therefore optimising the design for locating the NOEC is not a possibility.

(4) Basic statistical experimental practice should be followed.

This entails randomising the allocation of organisms to test concentrations and randomising the tanks or vessels in the room where the experiment is to be conducted. The use of blocking should be considered to control the effects of local environmental factors, for example, lighting and temperature gradients across a room. Blocking should be taken into account in subsequent statistical analysis.

(5) The numbers of replicates and individuals needed should calculated by considering the power of the statistical tests.

This requires a specification of what size of effect it is desired to detect and with what degree of certainty one wants to be able to detect it.

The expected level of experimental variation also needs to be known, for example from past experiments, or estimated in some way. Using these pieces of information it is then possible to calculate the approximate numbers required. The actual power can be determined after the experiment when the variability is known.

Such calculations were performed for the Fish Juvenile Growth Test. A table of possible designs, indicating their sensitivity, can be found as an appendix to the draft Guideline.

EC point estimation

(1) Design may be more critical for threshold models than for standard concentration-response curves.

For threshold models there should be some attempt to have sufficient concentrations around the perceived region of the threshold to enable a reasonable precision to be obtained.

Depending on the form of the concentration-response part of the model, then an adequate number of concentrations also needs to be used above the threshold to characterise the relationship.

Therefore, greater resources are probably needed, relative to EC*x* estimation.

(2) Optimal design for non-parametric analyses has concentrated on sequential experiments.

These are tests conducted one concentration at a time. Previous results in the series are used to determine the next test level. Experiments can be carried on until a given precision is obtained for the EC*x*, for example. Sequential experimentation can be very efficient in terms of experimental resources but may be unrealistic for aquatic toxicology experiments.

(3) Optimal designs can be determined for regression models.

Morgan (1992) provides a good review for quantal data and much of the literature can be extended to non-quantal situations. Silvey (1980) gives a statistical introduction to this area. Most work has assumed that the parameters of the model are known in advance which is somewhat unrealistic.

Some authors have attempted to incorporate uncertainty and derive optimal designs that are robust to possible variation in the model's location and slope. Chaloner and Larntz (1989) consider a uniform distribution for the parameters and choose to minimise, with respect to the design, the expected value of some function of the estimated variances. Kalish (1990) and Muller and Schmitt (1990) present other approaches, although these are possibly more restrictive in terms of the permissible designs.

It is not surprising that the better the prior knowledge an experimenter has, the smaller the experiment needed to estimate an EC*x*. This experiment is concentrated in the area where the EC*x* lies. Such an experiment might only comprise three or four concentrations and so be less costly than current Guidelines would require. Where there is little prior knowledge then, again, it is not surprising that more concentrations would be needed to make the experiment robust to the location of the unknown concentration-response curve.

To summarise, it should be possible to provide experimenters with help on optimal choice, and number, of concentrations given estimates or predictions of the likely ranges of parameters. This will go some way to producing more cost-effective designs that may actually require fewer resources than at present.

3.5 Conclusions

No method of analysis can be expected to be optimal in all circumstances. The NOEC has severe limitations as a summary statistic. EC point estimation generally has conceptual advantages and may lead to cost reductions in experiment but modelling data may have a number of practical drawbacks. Non-parametric methods of EC point estimation may lack flexibility.

EC point estimation via data modelling is therefore the preferred option because there are only practical objections to overcome, not problems related to interpretation or generality.

It is important to involve ecotoxicologists, regulators and those involved in risk assessment in the debate since recommendations for changing statistical analyses may have wide-ranging implications. One key issue is what level of effect is it biologically important to be able to estimate and with what precision.

More work is required to further explore the use of threshold models and the optimisation of experimental design for EC*x* estimation. Available data should be collected and re-examined in order that guideline-specific recommendations can be made on design and analysis.

4. Recommendations

(1) EC point estimation is recommended in preference to NOEC estimation.

The latter has a number of drawbacks that make it undesirable as a summary measure of toxic effect. EC point estimation may lead to more cost-effective experiments.

Consideration therefore needs to be given to the implications of changing from NOECs to EC*x*s for the current and draft Guidelines and for risk assessment.

(2) Member countries should consider extending their network of experts to include a list of statistical experts.

Statistical advice would then be available on present and future Guidelines. Workshops could be arranged to address general or specific topics of relevance to the Guidelines.

In particular the re-analysis of ring-test and other published data could be undertaken with a view to making definitive statements about recommended analyses for each Guideline.

(3) That OECD should promote the development of a handbook of statistical methods and computer software.

A handbook of statistical methods could be produced covering the techniques recommended in the Guidelines. This might be a practical alternative to supplying a large amount of statistical detail in individual Guidelines, especially as many of the Guidelines have very similar requirements.

Computer software is another area where there is likely to be increased interest in the future. Standard software would improve the consistency of reporting and greatly assist laboratories without access to professional statisticians. Any package should ideally contain guidance on experimental design.

5. Acknowledgements

I would like to thank Herman Koëter and Nicola Grandy of OECD for their help and support in the preparation of this report.

I would also like to thank Richard Stephenson and Ann Gould of Shell Research and Peter Chapman (Zeneca, U.K.) and Bob Lacey (Water Research Council, U.K.) for many constructive comments.

6. References

Barnthouse, L.W., Suter II, G.W., Rosen, A.E. and Beauchamp, J.J (1987). Estimating responses of fish populations to toxic contaminants. *Environ. Toxicol. Chem.*, **6**, 811-824.

Bruce, R.D. and Versteeg, D.J. (1992). A statistical procedure for modelling continuous toxicity data. *Environ. Toxicol. Chem.*, **11**, 1485-1494.

Buxton, D. (1982). Instructions to authors. *Agronomy J.*, **74**, 1100-1101.

Capizzi, T., Oppenheimer, L., Mehta, H., Naimie, H. and Fair, J.L. (1985). Statistical considerations in the evaluation of chronic aquatic toxicity studies. *Environ. Sci. Technol.*, **19**, 35-43.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *J. Stat. Plan. Inf.*, **21**, 191-208.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

Cousens, R. and Marshall, C. (1987). Dangers in testing statistical hypotheses. *Ann. Appl. Biol*, **111**, 469-476.

Cox, C. (1987). Threshold dose-response models in toxicology. *Biometrics*, **43**, 525-535.

Cox, D.R. (1958). *Planning of Experiments*. Wiley, New York.

Crossland, N.O. and La Point, T.W. (1992). Editorial: The design of mesocosm experiments. *Environ. Toxicol. Chem*, **11**, 1-4.

Crump, K.S. (1984). A new method for determining allowable daily intakes. *Fundam. Appl. Toxicol*, **4**, 854-871.

DeGraeve, G.M., Cooney, J.D., Marsh, B.H., Pollock, T.L. and Reichenbach, N.G. (1992). Variability in the performance of the 7-d *Ceriodaphnia dubia* survival and reproduction test: an intra- and interlaboratory study. *Environ. Toxicol. Chem.*, **11**, 851-866.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Soc.*, **50**, 1096-1121.

Dunnett, C.W. (1964). New tables for multiple comparisons with contol. *Biometrics*, **20**, 482-491.

Feder, P.I. and Collins, W.J. (1982). Considerations in the design and analysis of chronic aquatic tests of toxicity. *Aquatic toxicology and hazard assessment: fifth conference, ASTM STP 766*, J.G. Pearson, R.B. Foster and W.E. Bishop, Eds., American Society for Testing and Materials, 1982, 32-68.

Finney, D.J. (1971). *Probit Analysis*. 3rd Edition. Cambridge University Press.

Finney, D.J. (1978). *Statistical Methods in Biological Assay*. Griffin.

Gaylor, D.W. (1989). Quantitative risk analysis for quantal reproductive and developmental effects. *Environ. Health Perspec*, **79**, 243-246.

Hamilton, M.A., Russo, R.C. and Thurston, R.V. (1977). Trimmed Spearman-Kärber method for estimating median lethal concentrations in toxicity bioassays. *Environ. Sci. Technol.*, **11**, 714-718.

Hoekstra, J.A. and Van Ewijk, P.H. (1993). Alternatives for the no-observed effect level. *Environ. Toxicol. Chem*, **12**, 187-194.

Kalish, (1990). Efficient design for estimation of median lethal dose and quantal dose-response curves. *Biometrics,* **46**, 737-748.

Leisenring, W. and Ryan, L. (1992). Statistical properties of the NOAEL. *Regul. Toxicol. Pharmacol.,* **15**, 161-171.

Liber, K., Kaushik, N.K., Solomon, K.R. and Carey, J.H. (1992). Experimental designs for aquatic mescosm studies: a comparison of the "ANOVA" and "regression" design for assessing the impact of tetrachlorophenol on zooplankton populations in limnocorrals. *Environ. Toxicol. Chem,* **11**, 61-77.

Masters, J.A., Lewis, M.A., Davidson, D.H. and Bruce, R.D. (1991). Validation of a four-day *Ceriodaphnia* toxicity test and statistical considerations in data analysis. *Environ. Toxicol. Chem*, **10**, 47-55.

Mayer, F., Mayer, K.S. and Ellersieck, M.R. (1986). Relation of survival to other endpoints in chronic toxicity tests with fish. *Environ. Toxicol. Chem*, **5**, 737-748.

McClave, J.T., Sullivan, J.H. and Pearson, J.G. (1981). Statistical analysis of fish chronic toxicity test data. *Aquatic toxicology and hazard assessment: fourth conference, ASTM STP 737,* D.R. Branson and K.L. Dickson, Eds., American Society for Testing and Materials, 1981, 359-376.

Morgan, B.J.T. (1992). *Analysis of quantal response data*. Chapman and Hall, London.

Muller, H-G. and Schmitt, T. (1990). Choice of number of doses for maximum likelihood estimation of the EC50 for quantal dose-response data. *Biometrics*, **46**, 117-129.

Nyholm, N., Sorensen, P.S., Kusk, K.O. and Christensen, E.R. (1992). Statistical treatment of data from microbial toxicity tests. *Environ. Toxicol. Chem*, **11**, 157-167.

Oris, J.T. and Bailer, A.J. (1993). Statistical analysis of the *Ceriodaphnia* toxicity test: sample size determination for reproductive effects. *Environ. Toxicol. Chem,* **12**, 85-90.

Oris, J.T., Winner, R.W. and Moore, M.V. (1991). A four-day survival and reproduction toxicity test for *Ceriodaphnia dubia*. *Environ. Toxicol. Chem,* **10**, 217-224.

Silvey, S.D. (1980). *Optimal Design*. Chapman and Hall, London.

Skalski, J.R. (1981). Statistical inconsistencies in the use of no-observed- effect levels in toxicity testing. *Aquatic toxicology and hazard assessment: fourth conference, ASTM STP 737,* D.R. Branson and K.L. Dickson, Eds., American Society for Testing and Materials, 1981, 377-387.

Sloof, W., Van Oers, J.A.M. and de Zwart, D. (1986). Margins of uncertainty in ecotoxicological hazard assessment. *Environ. Toxicol. Chem,* **5**, 841-852.

Stephan, C.E. and Rogers, J.W. (1985). Advantages of using regression to calculate results of chronic toxicity tests. *Aquatic toxicology and hazard assessment: eighth symposium, ASTM STP 891*, R.C. Bahner and D.J. Hansen, Eds., American Society for Testing and Materials, 1985, 328-338.

Suter II, G.W., Rosen, A.E., Linder, E. and Parkhurst, D.F. (1987). Endpoints for responses of fish to chronic toxic exposures. *Environ. Toxicol. Chem*, **6**, 793-809.

Van der Hoeven, N. (1991). LC_{50} estimates and their confidence intervals derived for tests with only one concentration with partial effect. *Wat. Res.*, **25**, 401-408.

Walsh, G.E., Deans, C.H. and McLaughlin, L.L. (1987). Comparison of the EC50s of algal toxicity tests calculated by four methods. *Environ. Toxicol. Chem*, **6**, 767-770.

Westlake, G.F., Sprague, J.B. and Rowe, D.W. (1983). Sublethal effects of treated liquid effluent from a petroleum refinery. v. reproduction of *Daphnia pulex* and overall evaluation. *Aquatic Toxicol.*, **4**, 327-339.

Williams, D.A. (1971). A test for differences between treatment means when several does levels are compared with a zero dose control. *Biometrics*, **27**, 103-117.

Williams, D.A. (1972). The comparison of several dose levels with a zero dose control. *Biometrics*, **28**, 519-531.

Williams, D.A. (1986). Interval estimation of the median lethal dose. *Biometrics*, **42**, 641-646.