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

01-Aug-2012

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

FISH TOXICITY TESTING FRAMEWORK

Series on Testing and Assessment

No. 171

JT03325150

Complete document available on OLIS in its original format

ENV/JM/MONO(2012)16
Unclassified

English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 171

FISH TOXICITY TESTING FRAMEWORK

IOMC

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris 2012

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, most recently revised 2009)*

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. 10, *Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data (1998)*

No. 11, *Detailed Review Paper on Aquatic Testing Methods for Pesticides and industrial Chemicals (1998)*

No. 12, *Detailed Review Document on Classification Systems for Germ Cell Mutagenicity in OECD Member Countries (1998)*

No. 13, *Detailed Review Document on Classification Systems for Sensitising Substances in OECD Member Countries 1998)*

No. 14, *Detailed Review Document on Classification Systems for Eye Irritation/Corrosion in OECD Member Countries (1998)*

No. 15, *Detailed Review Document on Classification Systems for Reproductive Toxicity in OECD Member Countries (1998)*

No. 16, *Detailed Review Document on Classification Systems for Skin Irritation/Corrosion in OECD Member Countries (1998)*

No. 17, *Environmental Exposure Assessment Strategies for Existing Industrial Chemicals in OECD Member Countries (1999)*

No. 18, *Report of the OECD Workshop on Improving the Use of Monitoring Data in the Exposure Assessment of Industrial Chemicals (2000)*

No. 19, *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (1999)*

No. 20, *Revised Draft Guidance Document for Neurotoxicity Testing (2004)*

No. 21, *Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals (2000)*

No. 22, *Guidance Document for the Performance of Out-door Monolith Lysimeter Studies (2000)*

No. 23, *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures (2000)*

No. 24, *Guidance Document on Acute Oral Toxicity Testing (2001)*

No. 25, *Detailed Review Document on Hazard Classification Systems for Specifics Target Organ Systemic Toxicity Repeated Exposure in OECD Member Countries (2001)*

No. 26, *Revised Analysis of Responses Received from Member Countries to the Questionnaire on Regulatory Acute Toxicity Data Needs (2001)*

No 27, *Guidance Document on the Use of the Harmonised System for the Classification of Chemicals which are Hazardous for the Aquatic Environment (2001)*

No 28, *Guidance Document for the Conduct of Skin Absorption Studies (2004)*

No 29, *Guidance Document on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (2001)*

No 30, *Detailed Review Document on Hazard Classification Systems for Mixtures (2001)*

No 31, *Detailed Review Paper on Non-Genotoxic Carcinogens Detection: The Performance of In-Vitro Cell Transformation Assays (2007)*

No. 32, *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (2000)*

No. 33, *Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures (2001)*

No. 34, *Guidance Document on the Development, Validation and Regulatory Acceptance of New and Updated Internationally Acceptable Test Methods in Hazard Assessment (2005)*

No. 35, *Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies (2002)*

No. 36, *Report of the OECD/UNEP Workshop on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long Range Transport in the Context of PBTS/POPS Assessment (2002)*

No. 37, *Detailed Review Document on Classification Systems for Substances which Pose an Aspiration Hazard (2002)*

No. 38, *Detailed Background Review of the Uterotrophic Assay Summary of the Available Literature in Support of the Project of the OECD Task Force on Endocrine Disrupters Testing and Assessment (EDTA) to Standardise and Validate the Uterotrophic Assay (2003)*

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (2009)*

No. 40, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures which Cause Respiratory Tract Irritation and Corrosion (2003)*

No. 41, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures which in Contact with Water Release Toxic Gases (2003)*

No. 42, *Guidance Document on Reporting Summary Information on Environmental, Occupational and Consumer Exposure (2003)*

No. 43, *Guidance Document on Mammalian Reproductive Toxicity Testing and Assessment (2008)*

No. 44, *Description of Selected Key Generic Terms Used in Chemical Hazard/Risk Assessment (2003)*

No. 45, *Guidance Document on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long-range Transport (2004)*

No. 46, *Detailed Review Paper on Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances (2004)*

No. 47, *Detailed Review Paper on Fish Screening Assays for the Detection of Endocrine Active Substances (2004)*

No. 48, *New Chemical Assessment Comparisons and Implications for Work Sharing (2004)*

No. 49, *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs (2004)*

No. 50, *Report of the OECD/IPCS Workshop on Toxicogenomics (2005)*

No. 51, *Approaches to Exposure Assessment in OECD Member Countries: Report from the Policy Dialogue on Exposure Assessment in June 2005 (2006)*

No. 52, *Comparison of Emission Estimation Methods Used in Pollutant Release and Transfer Registers (PRTRs) and Emission Scenario Documents (ESDs): Case Study of Pulp and Paper and Textile Sectors (2006)*

No. 53, *Guidance Document on Simulated Freshwater Lentic Field Tests (Outdoor Microcosms and Mesocosms) (2006)*

No. 54, *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (2006)*

No. 55, *Detailed Review Paper on Aquatic Arthropods in Life Cycle Toxicity Tests with an Emphasis on Developmental, Reproductive and Endocrine Disruptive Effects (2006)*

No. 56, *Guidance Document on the Breakdown of Organic Matter in Litter Bags (2006)*

No. 57, *Detailed Review Paper on Thyroid Hormone Disruption Assays (2006)*

No. 58, *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals (2006)*

No. 59, *Report of the Validation of the Updated Test Guideline 407: Repeat Dose 28-Day Oral Toxicity Study in Laboratory Rats (2006)*

No. 60, *Report of the Initial Work towards the Validation of the 21-Day Fish Screening Assay for the Detection of Endocrine Active Substances (Phase 1A) (2006)*

No. 61, *Report of the Validation of the 21-Day Fish Screening Assay for the Detection of Endocrine Active Substances (Phase 1B) (2006)*

No. 62, *Final OECD Report of the Initial Work towards the Validation of the Rat Hershberger Assay: Phase-1, Androgenic Response to Testosterone Propionate, and Anti-Androgenic Effects of Flutamide (2006)*

No. 63, *Guidance Document on the Definition of Residue (2006)*

No. 64, *Guidance Document on Overview of Residue Chemistry Studies (2006)*

No. 65, *OECD Report of the Initial Work towards the Validation of the Rodent Uterotrophic Assay - Phase 1 (2006)*

No. 66, *OECD Report of the Validation of the Rodent Uterotrophic Bioassay: Phase 2. Testing of Potent and Weak Oestrogen Agonists by Multiple Laboratories (2006)*

No. 67, *Additional Data Supporting the Test Guideline on the Uterotrophic Bioassay in rodents (2007)*

No. 68, *Summary Report of the Uterotrophic Bioassay Peer Review Panel, including Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow up of this Report (2006)*

No. 69, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models (2007)*

No. 70, *Report on the Preparation of GHS Implementation by the OECD Countries (2007)*

No. 71, *Guidance Document on the Uterotrophic Bioassay - Procedure to Test for Antioestrogenicity (2007)*

No. 72, *Guidance Document on Pesticide Residue Analytical Methods (2007)*

No. 73, *Report of the Validation of the Rat Hershberger Assay: Phase 3: Coded Testing of Androgen Agonists, Androgen Antagonists and Negative Reference Chemicals by Multiple Laboratories. Surgical Castrate Model Protocol (2007)*

No. 74, *Detailed Review Paper for Avian Two-generation Toxicity Testing (2007)*

No. 75, *Guidance Document on the Honey Bee (Apis Mellifera L.) Brood test Under Semi-field Conditions (2007)*

No. 76, *Final Report of the Validation of the Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances: Phase 1 - Optimisation of the Test Protocol (2007)*

No. 77, *Final Report of the Validation of the Amphibian Metamorphosis Assay: Phase 2 - Multi-chemical Interlaboratory Study (2007)*

No. 78, *Final Report of the Validation of the 21-day Fish Screening Assay for the Detection of Endocrine Active Substances, Phase 2: Testing Negative Substances (2007)*

No. 79, *Validation Report of the Full Life-cycle Test with the Harpacticoid Copepods Nitocra Spinipes and Amphiascus Tenuiremis and the Calanoid Copepod Acartia Tonsa - Phase 1 (2007)*

No. 80, *Guidance on Grouping of Chemicals (2007)*

No. 81, *Summary Report of the Validation Peer Review for the Updated Test Guideline 407, and Agreement of the Working Group of National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2007)*

No. 82, *Guidance Document on Amphibian Thyroid Histology (2007)*

No. 83, *Summary Report of the Peer Review Panel on the Stably Transfected Transcriptional Activation Assay for Detecting Estrogenic Activity of Chemicals, and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2007)*

No. 84, *Report on the Workshop on the Application of the GHS Classification Criteria to HPV Chemicals, 5-6 July, Bern Switzerland (2007)*

No. 85, *Report of the Validation Peer Review for the Hershberger Bioassay, and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2007)*

No. 86, *Report of the OECD Validation of the Rodent Hershberger Bioassay: Phase 2: Testing of Androgen Agonists, Androgen Antagonists and a 5 α-Reductase Inhibitor in Dose Response Studies by Multiple Laboratories (2008)*

No. 87, *Report of the Ring Test and Statistical Analysis of Performance of the Guidance on Transformation/Dissolution of*

Metals and Metal Compounds in Aqueous Media (Transformation/ Dissolution Protocol) (2008)

No. 88, *Workshop on Integrated Approaches to Testing and Assessment (2008)*

No. 89, *Retrospective Performance Assessment of the Test Guideline 426 on Developmental Neurotoxicity (2008)*

No.90, *Background Review Document on the Rodent Hershberger Bioassay (2008)*

No. 91, *Report of the Validation of the Amphibian Metamorphosis Assay (Phase 3) (2008)*

No. 92, *Report of the Validation Peer Review for the Amphibian Metamorphosis Assay and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2008)*

No. 93, *Report of the Validation of an Enhancement of OECD TG 211: Daphnia Magna Reproduction Test (2008)*

No. 94, *Report of the Validation Peer Review for the 21-Day Fish Endocrine Screening Assay and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-Up of this Report (2008)*

No. 95, *Detailed Review Paper on Fish Life-Cycle Tests (2008)*

No. 96, *Guidance Document on Magnitude of Pesticide Residues in Processed Commodities (2008)*

No. 97, *Detailed Review Paper on the use of Metabolising Systems for In Vitro Testing of Endocrine Disruptors (2008)*

No. 98, *Considerations Regarding Applicability of the Guidance on Transformation/Dissolution of Metals Compounds in Aqueous Media (Transformation/Dissolution Protocol) (2008)*

No. 99, *Comparison between OECD Test Guidelines and ISO Standards in the Areas of Ecotoxicology and Health Effects (2008)*

No. 100, *Report of the Second Survey on Available Omics Tools (2009)*

No. 101, *Report of the Workshop on Structural Alerts for the OECD (Q)SAR Application Toolbox, 15-16 May 2008, Utrecht, the Netherlands (2009)*

No. 102, *Guidance Document for using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories According to the OECD Guidance on Grouping of Chemicals (2009)*

No. 103, *Detailed Review Paper on Transgenic Rodent Mutation Assays (2009)*

No. 104, *Performance Assessment: Comparsion of 403 and CxT Protocols via Simulation and for Selected Real Data Sets (2009)*

No. 105, *Report on Biostatistical Performance Assessment of the draft TG 436: Acute Toxic Class Testing Method for Acute Inhalation Toxicity (2009)*

No. 106, *Guidance Document for Histologic Evaluation of Endocrine and Reproductive Test in Rodents (2009)*

No. 107, *Preservative Treated Wood to the Environment for Wood Held in Storage after Treatment and for Wooden Commodities that are not Covered and are not in Contact with Ground. (2009)*

No. 108, *Report of the Validation of the Hershberger Bioassay (weanling model) (2009)*

No. 109, *Literature Review on the 21-Day Fish Assay and the Fish Short-Term Reproduction Assay (2009)*

No. 110, *Report of the Validation Peer Review for the Weanling Hershberger Bioassay and Agreement of the Working Group of National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2009)*

No. 111, *Report of the Expert Consultation to Evaluate an Estrogen Receptor Binding Affinity Model for Hazard Identification (2009)*

No. 112, *The 2007 OECD List of High Production Volume Chemicals (2009)*

No. 113, *Report of the Focus Session on Current and Forthcoming Approaches for Chemical Safety and Animal Welfare (2010)*

No. 114, *Performance Assessment of Different Cytotoxic and Cytostatic Measures for the In Vitro Micronucleus Test (MNVIT): Summary of results in the collaborative trial (2010)*

No. 115, *Guidance Document on the Weanling Hershberger Bioassay in Rats: A Short-term Screening Assay for (Anti) Androgenic Properties (2009)*

No. 116, *Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies, Supporting Test Guidelines 451, 452 and 453 – 2nd Edition (2011)*

No. 117, *Guidance Document 117 on the Current Implementation of Internal Triggers in Test Guideline 443 for an Extended One Generation Reproductive Toxicity Study, in the United States and Canada (2011)*

No. 118, *Workshop Report on OECD Countries' Activities Regarding Testing, Assessment and Management of Endocrine Disrupters Part I and Part II (2010)*

No. 119, *Classification and Labelling of chemicals according to the UN Globally Harmonized System: Outcome of the Analysis of Classification of Selected Chemicals Listed in Annex III of the Rotterdam Convention (2010)*

No. 120, *Part 1: Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic Chemistry Mechanism-based Structural Alerts for the Identification of DNA Binding Chemicals (2010)*

No. 120, *Part 2: Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic Chemistry Mechanism-based Structural Alerts for the Identification of DNA Binding Chemicals (2010)*

No. 121, *Detailed Review Paper (DRP) on Molluscs Life-cycle Toxicity Testing (2010)*

No. 122, *Guidance Document on the Determination of the Toxicity of a Test Chemical to the Dung Beetle Aphodius Constans (2010)*

No. 123, *Guidance Document on the Diagnosis of Endocrine-related Histopathology in Fish Gonads (2010)*

No. 124, *Guidance for the Derivation of an Acute Reference Dose (2010)*

No. 125, *Guidance Document on Histopathology for Inhalation Toxicity Studies, Supporting TG 412 (Subacute Inhalation Toxicity: 28-Day) and TG 413 (Subchronic Inhalation Toxicity: 90-Day) (2010)*

No. 126, *Short Guidance on the Threshold Approach for Acute Fish Toxicity (2010)*

No. 127, *Peer Review Report of the Validation of the 21-Day Androgenised Female Stickleback Screening Assay (2010)*

No. 128, *Validation Report of the 21-Day Androgenised Female Stickleback Screening Assay (2010)*

No. 129, *Guidance Document on Using Cytotoxicity Tests to Estimate Starting Doses for Acute Oral Systemic Toxicity Tests(2010)*

No. 131, *Report of the Test Method Validation of Avian Acute Oral Toxicity Test (OECD Test Guideline 223) (2010)*

No. 132, *Report of the Multi-Laboratory Validation of the H295R Steroidogenesis Assay to Identify Modulators (2010)*

No.133, *Peer Review Report for the H295R Cell-Based Assay for Steroidogenesis (2010)*

No.134, *Report of the Validation of a Soil Bioaccumulation Test with Terrestrial Oligochaetes by an International ring test (2010)*

No.135, *Detailed Review Paper on Environmental Endocrine Disruptor Screening: The use of Estrogen and Androgen Receptor Binding and Transactivation Assays in Fish (2010)*

No. 136, *Validation Report of the Chironomid Full Life-Cycle Toxicity Test (2010)*

No. 137, *Explanatory Background Document to the OECD Test Guideline on In Vitro Skin Irritation Testing (2010)*

No. 138, *Report of the Workshop on Using Mechanistic Information in Forming Chemical Categories (2011)*

No. 139, *Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic Chemistry Mechanism Based Structural Alerts for the Identification of Protein-binding Chemicals (2011)*

No. 140, *Report of the WHO/OECD/ILSI (Hesi) Workshop on Risk Assessment of Combined Exposures to Multiple Chemicals (2011)*

No. 141, *Report of the Phase 1 of the Validation of the Fish Sexual Development Test for the Detection of Endocrine Active Substances (2011)*

No. 142, *Report of the Phase 2 of the Validation of the Fish Sexual Development Test for the Detection of Endocrine Active Substances (2011)*

No. 143, *Peer Review Report for the Validation of the Fish Sexual Development Test and Agreement of the Working Group of National*

Co-ordinators of the Test Guideline Programme on the Follow-up of the Peer Review (2011)

No. 144, *Validation Report for the Acute Chironomid Assay (2011)*

No. 145, *Transgenic Rodent Somatic and Germ Cell Gene Mutation Assay: Retrospective Performance Assessment (2011)*

No. 146, *Syrian Hamster Embryonic (SHE) Cell PH 6.7 Cell Transformation Assay Prevalidation Study Report (2012)*

No. 147, *Syrian Hamster Embryonic (SHE) Cell PH 7.0 Cell Transformation Assay Prevalidation Study Report (2012)*

No. 148, *Guidance Document on the Androgenised Female Stickleback Screen (2011)*

No. 149, *Validation Report of the Balb/c 3T3 Cell Transformation Assay (2012)*

No. 152, *Case Study: Assessment of an Extended Chemical Category, the Short-chain Methacrylates, Targeted on Bioaccumulation (2011)*

No. 153, *Guidance Document for the Derivation of an Acute Reference Concentration (Arfc) (2011)*

No. 154, *Validation Report: Part 1 – Validation of Efficacy Methods for Antimicrobials used on Hard Surfaces (2011)*

No. 154, *Validation Report: Part 2 – Validation of Efficacy Methods for Antimicrobials used on Hard Surfaces (2011)*

No. 155, *Peer Review for the Validation of the Modified Skin Irritation Test Method using LabyCyte EPI-MODEL24; Additional Studies; and Agreement of the Working Group of National Coordinators on the Follow-up to the Peer Review* (2011)

No. 156, *Guidance Notes on Dermal Absorption* (2011)

No. 157, *Validation Report Phase 1 for the Zebrafish Embryo Toxicity Test (2011)*

No. 158, *Report of Progress on the Interlaboratory Validation of the OECD Harpacticoid Copepod Development and Reproduction Test (2011)*

No. 159, *Validation Report for the Skin Irritation Test Method using Labcyte Epi-Model24 (2011)*

No. 160, *Guidance Document on the Bovine Corneal Opacity and Permeability (BCOP) and Isolated Chicken Eye (ICE) Test Methods: Collection of Tissues for Histological Evaluation and Collection of Data on Non-Severe Irritants (2011)*

No. 161, *Peer Review Report for the Validation of the Stably Transfected Transcriptional Activation Assay for the Detection of Androgenic and Anti-Androgenic Activity of Chemicals (2011)*

No. 165, *Guidance Document on Crop Field Trials (2011)*

No. 166, *SIDS Initial Assessment Profiles agreed in the course of the OECD HPV Chemicals Programme from 1993 to 2013 (2012)*

No. 167, *Crosswalk of Harmonized U.S. - Canada Industrial Function and Consumer and Commercial Product Categories with EU Chemical Product and Article Categories (2012)*

No. 168, *The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins Binding to Proteins (2012)*

No. 169, *Use of the AOP to Develop Chemical Categories and Integrated Assessment and Testing Approaches (2012)*

No. 170, *Guidance Document for Demonstrating Efficacy of Pool and Spa Disinfectants and Field Testing*

ABOUT THE OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 34 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise 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 specialised committees and working 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 working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in ten different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides and Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/ehs/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organisations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 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. The Participating Organisations are FAO, ILO, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. UNDP is an observer. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, 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 this and many other Environment, Health and Safety publications, consult the OECD's World Wide Web site (www.oecd.org/ehs/)

or contact:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This document presents a review of fish toxicity testing for the regulatory purpose of chemical safety. The main focus is on fish toxicity, but fish bioaccumulation is also considered where relevant. The document was initially elaborated by a group of experts and reviewed at an OECD Workshop on a Fish Toxicity Testing Framework, held on 28-30 September 2010 in the United Kingdom. A review of regulatory needs for fish tests under various jurisdictions in OECD countries is provided in Chapter 2, followed by a review of statistical issues and general test considerations in Chapters 3 and 4, respectively. The document examines animal welfare concerns and alternatives to fish tests in Chapter 5. Chapter 6 provides a systematic review of existing and draft OECD Guidelines which use fish for toxicity or bioaccumulation studies. Finally Chapter 7 describes a generic framework for assessing the environmental hazards of chemicals using fish tests in the most efficient way. An Annex contains conclusions and recommendations made and agreed at the workshop in September 2010. The recommendations concern, among other aspects, possible improvements to existing Test Guidelines, development of guidance on specific issues, harmonisation of existing Test Guidelines for common issues, development of new Test Guidelines, and proposals for deletion of outdated Test Guidelines.

The project for developing this document was proposed by the United States in 2008. Comments were requested from the Working Group of the National Coordinators to the Test Guidelines Programme (WNT) in 2010 and 2011, and an expert meeting was held in September 2011. The WNT approved the document in April 2012, and the Joint Meeting of the Chemicals Committee and working Party on Chemicals, Pesticides and Biotechnology (Joint Meeting) agreed to its declassification on 26 July 2012.

This document is published under the responsibility of the Joint Meeting.

TABLE OF CONTENTS

ACRONYMS

3 "Rs" principle:	Reduction, Refinement, Replacement of animal tests
ADME:	Absorption, Distribution, Metabolism, Excretion
AFSS:	Androgenised Female Stickleback Screen
ANOVA:	Analysis of Variance
ASTM:	American Society for Testing and Materials
BAF:	Bioaccumulation Factor
BCF:	Bioconcentration Factor
CLP:	Classification, Labelling and Packaging (European Regulation on)
DHT:	Di-hydrotestosterone
DT_{50}:	Disappearance Time 50% (time after which 50% of the substance disappeared)
EC_x:	Concentration at which x% of the effect observed is measured
ECETOC:	European Centre for Ecotoxicology and Toxicology of Chemicals
ECHA:	European Chemicals Agency
ECVAM:	European Centre for the Validation of Alternative Methods
EDSP:	Endocrine Disrupters Screening Programme
EDTA:	Endocrine Disrupters Testing and Assessment
ELISA:	Enzyme-Linked Immuno Sorbent Assay
ELS:	Early Life-Stage
FET (or ZFET):	Fish Embryo Toxicity (Zebrafish Embryo Toxicity)
FFLCT (or FLCT):	Fish Full Life-Cycle Test
FIFRA:	Federal Insecticide, Fungicide and Rodenticide Act
FLC:	Fish Life-Cycle
FSDT:	Fish Sexual Development Test
GHS:	Globally Harmonised System (for Classification and Labelling)
GSI:	Gonado-Somatic Index
HPG:	Hypothalamo-Pituitary-Gonadal Axis
HPT	Hypothalamo-Pituitary-Thyroid Axis
HTS:	High Throughput Screen
ICATM:	International Cooperation on Alternative Toxicological Methods
ICCVAM:	Interagency Coordinating Committee on the Validation of Alternative Methods
ICE:	Interspecies Correlation Estimation
ITS:	Intelligent Testing Strategy
LC_{50}:	Lethal Concentration for 50% of animals
LOEC:	Lowest Observed Effect Concentration
MAD:	Mutual Acceptance of Data
MMGT (or MMT):	Medaka Multi-Generation Test
MTC:	Maximum Tolerable Concentration
MOA:	Mode of Action
NCCT:	National Centre for Computational Toxicology

NICNAS:	National Industrial Chemicals Notification and Assessment Scheme
NOEC:	No Observed Effect Concentration
OECD:	Organisation for Economic Cooperation and Development
OPPTS:	Office of Prevention, Pesticides and Toxic Substances
PBT:	Persistent, Bioaccumalative and Toxic
PEC:	Predicted Exposure Concentration
PNEC:	Predicted No Effect Concentration
(Q)SAR:	(Quantitative) Structure-Activity Relationship
REACH:	Registration, Evaluation, Authorisation of Chemicals
SSD:	Species-Sensitivity Distribution
TG:	Test Guidelines
TG 203	Fish, Acute Toxicity Test (1992). http://www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en
TG 204	Fish, Prolonged Toxicity Test: 14 Day Study (1984). http://www.oecd-ilibrary.org/environment/test-no-204-fish-prolonged-toxicity-test-14-day-study_9789264069985-en
TG 210	Fish, Early Life-Stage Toxicity Test (1992). http://www.oecd-ilibrary.org/environment/test-no-210-fish-early-life-stage-toxicity-test_9789264070103-en
TG 212	Fish, Short-Term Toxicity Test on Embryo and Sac-Fry Stages (1998). http://www.oecd-ilibrary.org/environment/test-no-212-fish-short-term-toxicity-test-on-embryo-and-sac-fry-stages_9789264070141-en
TG 215	Fish, Juvenile Growth Test (2000). http://www.oecd-ilibrary.org/environment/test-no-215-fish-juvenile-growth-test_9789264070202-en
TG 229	Fish Short Term Reproduction Assay (2009). http://www.oecd-ilibrary.org/environment/test-no-229-fish-short-term-reproduction-assay_9789264076211-en
TG 230	21-Day Fish Assay (2009). http://www.oecd-ilibrary.org/environment/test-no-230-21-day-fish-assay_9789264076228-en
TG 234	Fish Sexual Development Test (2011). http://www.oecd-ilibrary.org/environment/test-no-234-fish-sexual-development-test_9789264122369-en
USEPA:	United States Environmental Protection Agency
VTG:	Vitellogenin

1. INTRODUCTION

1. In 2008, the United States submitted a project proposal to the Working Group of the National Coordinators of the Test Guidelines Programme to develop a framework for fish toxicity testing. The project was intended to complete a comprehensive review of regulatory needs/data requirements for fish testing and review the currency of existing OECD fish Test Guidelines. Many OECD Test Guidelines were developed several decades ago and it was worth reconsidering their applicability to current regulatory requirements and to possible future developments. The ultimate output of the project was to develop a guidance document on a Fish Toxicity Testing Framework including recommendations for deleting/ adding/ updating OECD fish toxicity-related Test Guidelines, and to propose possible toxicity testing strategies to minimize fish toxicity testing in accordance with the 3Rs- principle (see definition in the OECD Guidance Document No. 19 (OECD, 2000). Initially, a steering group of experts was formed to plan and develop materials for a larger expert workshop to address status of existing and proposed fish Test Guidelines, emerging testing needs, existing testing frameworks, and to propose a harmonized hazard testing scheme.

2. The objectives of the Fish Toxicity Testing Framework document are to: 1) provide an overview of existing fish toxicity Test Guidelines or Test Guidelines in preparation with the view to suggest developing new Guidelines, or updating or deleting existing Test Guidelines, and 2) suggest possible fish toxicity testing strategies with a view to minimizing fish toxicity testing. The document also reviews a range of general considerations for fish testing, including statistical aspects of test design and data analysis, and reviews animal welfare considerations. The Fish Toxicity Testing Framework document ends with a chapter on conclusions and recommendations for possible future work. Test Guidelines referred to in this document are those which had been adopted by October 2011 or earlier. Furthermore, it should be noted that the document deals primarily with fish toxicity testing, and goes into less detail for bioconcentration testing.

3. It should be noted that whenever explanatory text is provided in this document in relation to assessing chemicals for their endocrine disrupting properties, the reader should seek additional guidance on these aspects directly in the OECD Guidance Document on the Assessment of Chemicals for Endocrine Disruption (OECD, 2011). The Guidance Document goes in-depth into analysis of test results and information available across the board for a weight-of-evidence determination of endocrine properties of tested chemicals, which is not the purpose of the fish toxicity testing framework presented here.

References

4. OECD (2000), Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation, Series on Testing and Assessment No.19, ENV/JM/MONO(2000)7, OECD, Paris

5. OECD (2011), Guidance Document on the Assessment of Chemicals for Endocrine Disruption, Series on Testing and Assessment (draft), OECD, Paris

2. REGULATORY NEEDS AND DATA REQUIREMENTS FOR FISH TESTING

6. Perhaps the most important point to be made is that the fish tests which are the subject of OECD test guidelines (TG) have all been required for use in regulatory testing at one time or another in various jurisdictions. In other words, there are no guidelines for which there has been no perceived regulatory need somewhere (see Table 2.1). While recognizing this need, it should nevertheless be highlighted that some tests are used much more frequently or more widely than others. Furthermore, three of these (OECD TG 229, the Fish Short Term Reproduction Assay; OECD TG 230, the 21 d Fish Screening Assay; and TG 234, the Fish Sexual Development Test) were only published in 2009-2011 for use in endocrine disrupter (ED) screening and testing, and have therefore not yet been widely applied in routine testing programmes or according to legal requirements, although draft versions of them have been in informal use for several years. A version of OECD TG 229 (OPPTS 890.1350) is already in use in Tier 1 of the USEPA Endocrine Disruptor Screening Programme.

7. The point about regulatory need is hardly surprising given that all OECD test guidelines were originally developed at the request of OECD member countries. Prioritization for test guideline development includes consideration not only of whether the method is mature for standardization, but also whether there is a regulatory need. However, as many of the test guidelines were published several decades ago, it is time to reconsider their applicability to modern regulatory requirements and to possible future developments.

8. Acute fish toxicity data are widely required by regulatory authorities, even though such tests are less ethically acceptable than tests with plants or most invertebrates. Regulations may require fish toxicity data for three main reasons: First, fish toxicity data are often used, together with invertebrate and algae toxicity data, for hazard classification and labelling of chemicals (cf. section 2.1). Second, regulatory authorities need to know whether a substance or discharge is likely to cause fish mortalities (cf. section 2.2). Third, as long-term aquatic toxicity data are often lacking, acute fish toxicity data are often used together with short-term data on other pelagic species, such as algae and daphnia, for extrapolating to a predicted chronic no-effect concentration (e.g. PNEC for aquatic organisms) through the use of assessment factors (cf. section 2.3). It is generally acknowledged that use of such assessment factors for extrapolating to long-term concentrations which are safe for the environmental compartment of concern implies uncertainty. This uncertainty can conceptually be separated into several parts when based on short-term data. Sources of uncertainty include extrapolation:

- from short-term to long-term toxicity;
- from effect-concentrations on a few laboratory toxicity test species to effects on all species in the environmental compartment of concern;
- from single species direct toxic effects in the laboratory under constant conditions to multispecies effects (including indirect effects affecting interactions between species) in a temporally and spatially varying environment;
- from laboratory to laboratory, and from time to time within laboratories (inter- and intra-laboratory variation).

9. In order to gain ecologically relevant data on processes such as survival, growth, development or reproduction, regulations also require various types of long-term toxicity data on pelagic species including fish, typically either for certain use categories of chemicals regulated by

authorisation which cause exposure of the aquatic environment (e.g. certain pesticides, biocides, drugs and veterinary medicines) or for industrial chemicals but at higher assessment tiers, in order to minimize the uncertainty of the PNEC estimation. In this case, assessment factors are smaller than when considering acute data alone. In addition, species sensitivity distributions can be used as an alternative to the application of assessment factors, if sufficient amounts and diversity of high quality data are available (e.g. Maltby *et al*. 2005). The ultimate goal when establishing a PNEC is to protect populations of potentially more sensitive species living in the relevant environmental compartment from long-term decline. Hence, long-term pelagic toxicity tests on fish include apical data on survival or mortality, growth, development and/or reproduction after long-term exposure or exposure during sensitive life-stages. In some circumstances, there is a desire to add mechanistic endpoints to certain long-term tests in order to help explain a chemical's mode of action. An example would be the addition of vitellogenin induction to a fish full life-cycle test that is being used to evaluate a potential endocrine active chemical. This may be valuable, if mechanistic data are unavailable and if cause-effect relationships require clarification; this may apply especially in jurisdictions which seek to regulate substances on the basis of their intrinsic hazards, such as potential endocrine activity, rather than based on their environmental risks which are normally estimated by a comparison of effect levels with predicted exposure levels. However, such additions are not needed routinely for the majority of substances, for which apical data alone are regarded as sufficient.

10. A further reason for employing testing of fish is that regulatory authorities need to be confident that substances will not bioaccumulate in fish tissues to levels which may harm fish themselves or their consumers (either humans or wildlife). Negligible bioaccumulation can for most chemicals be confidently predicted from physical-chemical data (when the mechanism of action is driven by accumulation of body lipids). If the log octanol-water partition coefficient (K_{ow}) value is low, the bioconcentration factor (BCF) will also usually be low and the potential for bioaccumulation would be negligible. Conversely, a high K_{ow} value is indicative of a potentially high BCF in fish and may therefore be followed up with a fish bioconcentration test (OECD TG 305). This latter test guideline is currently being revised to include the possibility of reducing the cost and number of laboratory animals used, when this can be done without compromising the BCF determination. The revision also includes a possibility to estimate a bioaccumulation factor (BAF) from dietary exposure of the fish, when such a test design is warranted, because the high hydrophobicity of the substance implies difficulties in exposing the fish *via* water and because bioaccumulation by the dietary route may be of environmental relevance in itself. TG 305 and the current development of an updated TG 305 were not discussed at the workshop held in the United Kingdom, and is not discussed in other chapters of this document.

11. All OECD fish test guidelines address one or other of these objectives to gain insight into the acute or long-term toxicity or the bioaccumulative behaviour of chemicals in fish. Their ultimate purpose is to protect the long-term sustainability of aquatic species, including that of fish populations and fisheries, in the most efficient and ethically sound manner. This is accomplished by providing data useful for hazard and risk assessment for aquatic species, top predators of aquatic food chains, and secondary poisoning of humans indirectly exposed *via* aquatic food chains. Several OECD toxicity test guidelines are complementary for some endpoints and they therefore should be selected by considering their environmental and ethical justifications, together with the regulatory requirements in different jurisdictions. These matters are discussed further in Chapter 7.

Table 2.1: Fish testing requirements of various OECD jurisdictions

	OECD TG 203 (1992)	**OECD TG 204 (1984)**	**OECD TG 210 (1992)**	**OECD TG 212 (1998)**	**OECD TG 215 (2000)**	**OECD TG 229 (2009)**	**OECD TG 230 (2009)******	**OECD TG 234 (2011)**	**OECD TG 305 (1996)**
Title / Legislation	**Fish acute toxicity test**	**Fish prolonged toxicity test: 14-day study**	**Fish early life-stage toxicity test**	**Fish short-term toxicity test on embryo and sac-fry stages**	**Fish juvenile growth test**	**Fish short-term reproduction assay**	**21-Day fish screening assay**	**Fish sexual development test**	**Bioconcentration: Flow-through fish test**
EU Regulation (EC) No 1107/2009 on plant protection products *	Always for rainbow trout and warm water species. Revisions may allow for rainbow trout only. Formulation rainbow trout only	If not acutely toxic (> 0.1 mg/L acute LC_{50}); if Early Life Stage (ELS)/Full Life Cycle tests (FLC) are not appropriate – however, OECD TG 204 data have restricted relevance regards chronic toxicity (ELS/FLC tests preferable); only with combined with sublethal endpoints of	If BCF > 100 and/or LC_{50} < 0,1 mg/L and/or DT90(w/s) > 100. Generally not on formulation. [continued exposure]		If not acutely toxic (>0.1 mg/L acute LC_{50}); generally not on formulation; [continued exposure]; if ELS/FLC are not appropriate - however, OECD 215 data have restricted relevance regards chronic toxicity (ELS/FLC	*Ad hoc* basis, if concern for endocrine disruption	*Ad hoc* basis, if concern for endocrine disruption	Probably on an *ad hoc* basis, if concern for endocrine disruption	If log Kow > 3, and DT_{50} from water-sediment study >10d etc.

	OECD TG 203 (1992)	**OECD TG 204 (1984)**	**OECD TG 210 (1992)**	**OECD TG 212 (1998)**	**OECD TG 215 (2000)**	**OECD TG 229 (2009)**	**OECD TG 230 (2009)******	**OECD TG 234 (2011)**	**OECD TG 305 (1996)**
		TG 215 and exposure ≥21-days. Generally not on formulation			preferable)				
EU Regulation (EC) No 1107/2009 on plant protection products **	Always required for rainbow trout		Always required if exposure of surface water is likely and the compound is stable in water (<90% loss over 24 h *via* hydrolysis)			Should be conducted		Probably on an *ad hoc* basis, if concern for endocrine disruption	If log Kow > 3 and < 90% loss of original substance over 24 h *via* hydrolysis
US Federal Insecticide, Fungicide and Rodenticide Act (FIFRA)	Cold and warm water freshwater species and 1 saltwater fish species		Required in fresh-water species; con-ditionally required in saltwater species			Specifically under the EDSP			Conditionally required
EU Biocidal Products Directive (98/8/EC)	Base set requirement with one freshwater species (+	Usually not relevant for risk assessment	PNEC refinement	PNEC refinement (if log Kow < 4)	PNEC refinement (if log Kow < 5).	*Ad hoc* basis, if concern for endocrine disruption	*Ad hoc* basis, if concern for endocrine	Probably on an *ad hoc* basis, if concern for endocrine	Required for anti-foulings; if log Kow ≥ 3 or detergents (surface

	OECD TG 203 (1992)	**OECD TG 204 (1984)**	**OECD TG 210 (1992)**	**OECD TG 212 (1998)**	**OECD TG 215 (2000)**	**OECD TG 229 (2009)**	**OECD TG 230 (2009)******	**OECD TG 234 (2011)**	**OECD TG 305 (1996)**
	marine species, if relevant)						disruption	disruption	tension ≤ 50 mN/m)
EU industrial chemicals (REACH Regulation (EC) No. 1907/2006)***	If > 10 tonnes /year		If > 100 tonnes /year	If >100 tonnes /year	If > 100 tonnes /year	*Ad hoc* basis, if concern for endocrine disruption	*Ad hoc* basis, if concern for endocrine disruption	Probably on an *ad hoc* basis, if concern for endocrine disruption	If > 100 tonnes /year (derogations possible)
US industrial chemicals (Toxic Substances Control Act (TSCA))	Conditional requirement		Conditional requirement			Conditional requirement			Conditional requirement
EU Veterinary Pharmaceuticals (Regulation EC 726/2004)	Base set requirement for Tier A		Requirement for Phase II Tier B			Conditional requirement	Conditional requirement	Probably on an *ad hoc* basis, if concern for endocrine disruption	Required in Tier B, if log Kow is ≥ 4
EU Human Pharmaceuticals (Regulation EC 726/2004)			Base set requirement for Phase II Tier A			Conditional requirement	Conditional requirement	Probably on an *ad hoc* basis, if concern for endocrine disruption	Required for PBT screening, if log Kow is ≥ 4.5 and in Tier B if log Kow is ≥ 3
US FDA-CDER	Required for all human								

	OECD TG 203 (1992)	**OECD TG 204 (1984)**	**OECD TG 210 (1992)**	**OECD TG 212 (1998)**	**OECD TG 215 (2000)**	**OECD TG 229 (2009)**	**OECD TG 230 (2009)******	**OECD TG 234 (2011)**	**OECD TG 305 (1996)**
(1998)	pharmaceuticals								
Australian Industrial Chemicals (Notification and Assessment) Act 1989	Base set requirement for new chemicals								
Canadian Plant Protection Product Active Substances (Pest Management Regulatory Agency)	Base set requirement for cold and warm water freshwater fish.		Conditional requirement						Conditional requirement if log Kow is ≥ 3
Japanese Chemical Substances Control Law	Base set requirement with one freshwater species among OECD recommended fishes		Conditional requirement						Base set requirement with one freshwater species among OECD recommended fishes
Japanese Agri-cultural Chemicals Regulation	Base set requirement with carp								Conditionally required

* According to SANCO/3268/2001 rev.4 (final) 17 October 2002, Working Document, Guidance document on Aquatic Ecotoxicology in the context of the Directive 91/414/EEC

**According to SANCO/11843/2010 (draft) rev. July 2010 (revision of the above guidance document)

*** Chronic fish tests are only required if the chemical safety assessment indicates the need.

**** It should be noted that a version of TG 230, the Androgenised Female Stickleback Screen (AFSS) (OECD GD 148), with high sensitivity to androgens and anti-androgens, was published as a Guidance Document in 2011 (OECD, 2011). In due course, this may also be used by some jurisdictions for screening suspected endocrine disrupters, and it is already being used on a research basis.

12. The only jurisdiction which at present requires fish tests for possible endocrine activity is the USEPA Endocrine Disruptor Screening Program (EDSP) (http://www.epa.gov/endo/), which has issued test orders for a battery of endocrine screening assays that include a version of TG 229. Further details of the US approach are given below, together with expected regulatory approaches in the European Union and Japan.

13. Of the fish-based tests which are currently being developed into OECD guidelines or guidance documents (the Dietary Fish Bioconcentration Test/reduced fish bioconcentration test; the Fish (zebrafish) Embryo Toxicity Test; and the Japanese medaka Multi-Generation Test), all are already used on an *ad hoc* basis for research or in relation to specific regulatory purposes.

14. The rest of this chapter reviews the fish testing requirements of a range of regulations in several OECD jurisdictions covering various types of chemicals (pesticides, biocides, industrial chemicals, pharmaceuticals). Table 2.1 lists those requirements.

2.1 Classification and Labelling of Chemicals

15. In several jurisdictions in OECD and non-OECD countries (at least in the more developed regions), the majority of substances are subject to classification, labelling, packaging and transportation regulations. These pieces of legislation often include requirements for certain restrictions, if the substance is deemed to be hazardous to the environment. In order to address, more uniformly, the potential hazard of chemicals, the United Nations have adopted a Globally Harmonised System (GHS) of Classification and Labelling of Chemicals (http://www.unece.org/trans/danger/publi/ghs/ghs_rev03/03files_e.html), which is in use in many regions around the world. The GHS addresses the classification of chemicals based on types of hazard with a goal of providing harmonized rules, regulations, and hazard communication, including labels and safety data sheets, targeted to enhance the protection of human health and the environment during handling, transportation and use of chemicals. The GHS itself does not include requirements for testing of substances and mixtures to establish hazard classes. Instead, the GHS exclusively makes use of existing data.

16. In the GHS, hazard categories are established based on physico-chemical properties, degradation and fate, as well as toxicity parameters. In the European Union (EU), for example, most substances and chemical products other than medicines, veterinary medicines, cosmetics and food additives have to be classified and hence labelled according to criteria set out by the implementation of GHS in the EU, i.e. according to the Regulation (EC) No 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures and a Guidance published by the European Chemicals Agency (2009) (http://guidance.echa.europa.eu/docs/guidance_document/clp_en.pdf). For the environment, the CLP criteria are based on the most sensitive species tested (algae, *Daphnia*, fish); however, like GHS, CLP *per se* does not require new testing on animals. In general, testing on animals should be avoided wherever possible, and alternative methods (including *in vitro* testing, the use of (Q)SAR, read-across and/or category approaches) must always be considered first provided they give adequate reliability and quality of data (i.e. weight-of-evidence approach). Similar regulations implementing GHS, with relatively minor differences, can be found *inter alia* in North American, Korean, Chinese, Australasian and Japanese regulations.

17. With regard to the use of fish in the GHS and other categorization schemes, one or more of the existing OECD TGs, including TG 203, TG 305, TG 210, TG 212 and TG 215 (plus toxicity data for aquatic invertbrates and algae/plants) can be used as a source of

data for the aquatic hazard classification of chemical substances according to the GHS criteria. These criteria address acute and chronic hazards depending on the results of toxicity studies, bioaccumulation potential/actual bioaccumulation and biodegradationn studies. Hazard categories developed from these results are associated with particular labelling and transport restrictions, including (in the EU) a special label pictogram, a hazard statement (e.g. Toxic to Aquatic Life), and various precautionary statements (e.g. Avoid Release to the Environment).

2.2 Hazard identification and risk assessment

18. Hazard identification is one essential precursor to risk assessment and hence is integral to most systems of chemical regulation. Hazard and risk assessment schemes are usually organized in a tiered or conditional fashion, and only progress to higher, more data-intensive tiers, if no-observed-effect concentrations (NOEC or EC_{10}) are less than some pre-determined hazard level, or if predicted environmental concentrations exceed predicted no effect concentrations (PEC/PNEC). In some regulatory contexts, however, identification of a particular hazard or mode of action (e.g. endocrine activity or persistent, bioaccumulative and toxic compounds (PBT)) may, or will in the future, be considered sufficient evidence of potential hazard or particular concern, irrespective of predicted risk expressed as PEC/PNEC ratio. Progression to higher tiers of assessment may also occur, when certain production/sales tonnage triggers per manufacturer or importer are exceeded. Many regulatory risk assessment schemes require data on toxicity to fish (as well as on crustaceans and algae), and most specify the use of OECD fish test guidelines or their close equivalents. Table 2.1 shows that under one condition or another, these schemes employ fish tests from the list of relevant OECD test guidelines.

19. The amount of fish toxicity data required for a particular chemical is driven by the particular use (and/or tonnage marketed per manufacturer or importer per year) of the substance. Thus, regulatory schemes for pesticides and biocides (e.g. the US Federal Insecticide, Fungicide and Rodenticide Act (FIFRA), the European Biocidal Products Directive 98/8/EC, the European Pesticide Regulation (EC) No 1107/2009) tend to require more fish toxicity or longer term fish toxicity-related data than schemes aimed at new and existing industrial chemicals (e.g. REACH (EU 2006), US Toxic Substance Control Act Regulations (US EPA 1976), Australian Industrial Chemicals (Notification and Assessment) Act 1989 (NICNAS 1990)) or pharmaceuticals and veterinary medicines (EMEA 2004, 2006; FDA-CDER, 1998). Note that under some of the less data-intensive regulatory schemes, it is, in certain cases, possible to accept reliable (Q)SARs model predictions on aquatic organisms including fish instead of measured short-term toxicity data.

2.2.1 Pesticides and biocides

20. The fish toxicity data requirements of the US Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) (http://ecfr.gpoaccess.gov/cgi/t/text/textidx?c=ecfr&sid=013b05537f6069487ae3f2252ae1d5a0&rgn=div5&view=text&node=40:23.0.1.1.9&idno=40#40:23.0.1.1.9.7.1.1), focused on the regulation of plant protection products and biocides, are quite similar to those imposed by the EU (EU 2002) and many other jurisdictions (e.g. Canadian Pest Management Regulatory Agency). In essence, they require an acute fish toxicity test (freshwater and estuarine/marine) equivalent to OECD TG 203 for most pesticide use patterns, as well as a freshwater fish early life stage (ELS) test (equivalent to OECD TG 210). Marine fish ELS data and fish life cycle test data (USEPA 850.1500, Benoit 1982) are conditional requirements dependent on likely exposure scenarios or expected bioconcentration, or on

alerts from earlier ELS testing. Fish bioconcentration data are also required conditionally depending on the octanol-water partition coefficient of the test substance.

21. The main difference of EU pesticides requirements (EU 2002) from the US pesticide legislation (USEPA 1972) and other jurisdictions is that short-term toxicity data from a 14 day prolonged study (OECD TG 204) may occasionally be requested as a supplement to (or in place of) OECD TG 203, but this is quite rare (and TG 204 will not be requested according to the upcoming guidance document (EU 2010)). More often, the EU accepts 'chronic' data on juvenile fish growth (OECD TG 215), or even data from the embryo and sac fry test (OECD TG 212) under the Biocidal Products Directive (EU 1998), although the ELS (OECD TG 210) is still the preferred method of predicting true chronic toxicity, and is generally considered more sensitive than both OECD TG 212 and TG 215. Review data are now available (Oris *et al*. 2012) which suggest that OECD TG 210 could be made even more sensitive by increasing the number of replicates per treatment. Like the growth test, the ELS test is not in itself a true chronic test, but it is generally regarded as a good predictor of the effects of systemically toxic chemicals in full life-cycle tests (McKim, 1977). As with FIFRA, if true chronic fish data are required, the Pesticides Regulation (EC) No. 1107/2009 (EU 2009) specifies a fish full life-cycle test according to the USEPA guideline (in the absence of an OECD TG for this test; USEPA 850.1500; Benoit, 1982).

2.2.2 Pharmaceuticals

22. Environmental testing of human pharmaceuticals has been required in the United States since the 1980s (FDA-CDER, 1998). Elsewhere, human pharmaceuticals have not, until recently, been subjected to environmental testing, although veterinary pharmaceuticals in the EU have been assessed for possible environmental effects since 1996. In the EU, the European Agency for the Evaluation of Medicinal Products (now the European Medicines Agency) requires a fish early life stage test (OECD TG 210) of human pharmaceuticals as a fundamental requirement (on the grounds that exposure in the environment is likely to be semi-continuous, so acute testing is deemed irrelevant; EMEA 2006). A fish bioconcentration test (OECD TG 305) is also conditionally required, if bioaccumulation is expected on grounds of having a high octanol-water partition coefficient. Chemicals with specific modes of action e.g. some endocrine active substances, require fish tests including both early life-stage and sexual development as well as reproduction, since the fish early life-stage test may not reflect the most sensitive life-stages and/or the most sensitive endpoints.

23. The testing requirements for veterinary pharmaceuticals in the EU are rather similar to those described above for human medicines, except that a fish acute test (OECD TG 203) is required as base-set data in Tier A, and conduct of tests in Tier B is conditional on unsatisfactory risk quotients derived in Tier A (VICH, 2003).

24. Pharmaceuticals are at present also tested for their environmental properties in Australia.

2.2.3 General (or industrial) chemicals

25. In the EU, the hazard testing of general chemicals is conducted under the Regulation (EC) No 1907/2006 of the European Parliament and the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) (EU 2006). This is a tiered testing system, with progression to higher tiers partly driven by the tonnage of substance per manufacturer or importer into the EU and partly by the outcome of the risk assessment. It may require *inter alia* an acute fish toxicity test (e.g. OECD TG 203), but only for substances produced in quantities above 10 tonnes/year. If tonnage exceeds 100

tonnes/year, more definitive fish tests may be required, if the chemical safety assessment indicates the need to further investigate effects on aquatic organisms. The need may be indicated by a PEC/PNEC ratio above 1, but also by information concerning high acute to chronic ratios of structural analogues or physical-chemical parameters indicating poor water solubility or a high bioconcentration potential. These tests include (as appropriate) the short term toxicity test on embryo and sac-fry stages (OECD TG 212), the fish juvenile growth test (OECD TG 215), or the early life stage test (OECD TG 210), although only the latter seems to be used to any great extent. Above 100 tonnes/year, the fish bioconcentration test (OECD TG 305), a fish dietary study, and/or bioaccumulation assessment of sediment-dwelling benthic oligochaetes (OECD 315) is/are required, if the substance is predicted to be bioaccumulative on the basis of its bioaccumulation potential (e.g. high octanol-water partition coefficient) and if existing and/or alternative data are not sufficient. This may also be the case in a definitive PBT assessment for substances produced or imported in quantities > 10 tonnes/year. Chronic aquatic toxicity testing[1], in addition, may be proposed by the registrant, when a substance has low water solubility and no acute toxicity is expected.

26. It should be noted that even though laboratory animal welfare is considered and implementation of the principles of the 3Rs are made under many regulations, REACH is one of the few pieces of chemicals regulation which makes an explicit attempt to minimise fish testing to an extent that is consistent with environmentally safe chemical use. This effort to reduce fish testing is very welcome from the ethical point of view, but care must be taken not to dilute the degree of environmental protection and to allow investigations using fish tests in case of environmental concern.

27. REACH makes it possible to require evaluation of possible endocrine active properties, when an EU Member State takes the lead in performing a substance evaluation.

28. In the US, testing of general chemicals is conducted under the provisions of the Toxic Substances Control Act (TSCA – 1976, http://www.epa.gov/compliance/civil/tsca/tscaenfstatreq.html). This does not contain tonnage triggers, but all available data on a substance are reviewed in a risk assessment before deciding which types of further testing might be needed. Test data on acute toxicity to fish (e.g. OECD TG 203) are sometimes required, while potential risk for the pelagic compartment and the expectation of long-term exposure generally triggers the fish ELS test (OECD TG 210) and/or the FFLCT (USEPA 850.1500, Benoit 1982).

29. Several other jurisdictions (e.g. Australia) use a similar range of fish tests, and risk assessments, for evaluating general (industrial) chemicals / chemicals in commerce.

2.2.4 Possible endocrine activity assessment (including pesticides, biocides, pharmaceuticals and general chemicals etc.)

30. The only jurisdiction which at present explicitly requires fish tests for possible endocrine disrupters is the USEPA Endocrine Disruptor Screening Program (http://www.epa.gov/endo/), which in Phase 1 of the program is currently requiring the conduct of a battery of endocrine screening assays (Tier 1) on 67 specifically selected substances, most of which are pesticides. The substances were selected on the basis of exposure. These assays consist of 5 *in vitro* screens for receptor-mediated activity, and 6 *in vivo* screens. Of the latter, two involve screens with wildlife organisms (fish and

[1] If further testing on aquatic species is required for the toxicity (T) assessment after (v)P(v)B has been confirmed, testing would typically first target chronic *Daphnia* exposures.

amphibians), one of which is the fish short-term reproduction assay (OECD TG 229) conducted with fathead minnow (*Pimephales promelas*) and the other is the amphibian metamorphosis assay (OECD TG 231) with the African clawed frog (*Xenopus laevis*). Data from these Tier 1 screens are not yet available, but it is expected that OECD TG 229 will be able to detect estrogens, androgens, and aromatase inhibitors. OECD TG 229 may also be sensitive to some estrogen and androgen antagonists, although more data are required on this point. On the other hand, TG 231 is sensitive to thyroid disrupters. Note that Tier 1 tests are not designed to individually or definitively determine whether a substance is an endocrine disruptor acting on apical endpoints in wildlife or humans. Definitive testing requirements in Tier 2 of substances which are positive in Tier 1 have not yet been finalized by the USEPA, but the goals of Tier 2 testing for endocrine-related effects include, but are not necessarily limited to, a more conclusive determination of whether a substance has potential to cause apical effects through interaction with estrogen, androgen, and thyroid pathways at critical life stages, in a variety of taxonomic groups and through relevant exposure routes, and encompassing a broad range of doses. Tier 2 may potentially include either a fish life-cycle test (USEPA 850.1500, Benoit 1982) with additional endocrine-specific endpoints, or a Japanese medaka (*Oryzias latipes*) multi-generation test (OECD 2002). Inclusion, interpretation and use of validated endocrine-specific endpoints in these Tier 2 assays have not yet been well defined.

31. In the EU, ED has been recognised as a mode of action of concern. New EU legislation (EU 2009) now requires that pesticides should be assessed for their possible endocrine-active properties, but the precise criteria to be used have not yet been developed, agreed, adopted and published. Furthermore, endocrine disrupting properties of industrial chemicals also need to be considered under the REACH legislation, although the precise regulatory definition of EDs has not yet been finalised. However, it is assumed that the 21d fish screen (OECD TG 230), the androgenised female stickleback screen (OECD GD 148), the fish short-term reproduction assay (OECD TG 229), the fish sexual development test (OECD TG 234), an enhanced fish full life-cycle test (FFLCT) with endocrine-sensitive endpoints (e.g. a development of Benoit 1982; Länge et al., 2001; Wenzel et al., 2001) or the Japanese medaka multi-generation test (MMGT, OECD 2002) will be considered in this context. In the newly revised Plant Protection Regulation (EU 2009) and in the new Biocidal Products Regulation, approved in January 2012 by the European Parliament, provisions have been established to significantly minimise exposure to chemicals causing ED by stating that substances having ED properties cannot be approved as active substances, safeners or synergists if they are used in a way leading to any significant exposure.

32. It seems likely that other jurisdictions will follow suit, although none have yet enacted legislation precisely specifying information requirements regarding ED properties. For example, in Japan, the Ministry of the Environment has been studying various aspects of endocrine disruption since 1998, and has focused on chemicals found in the Japanese environment. In due course, it is expected to introduce *in vitro* and *in vivo* screening of certain chemicals, and in the case of fish the main candidate procedures are OECD TG 229 as the *in vivo* screen, and the medaka multi-generation test (MMGT) (or a similar multigeneration test with another species such as zebrafish) as the definitive test for use in decision-making. In the EU, industrial chemicals having ED properties have also been regarded as substances of very high concern. In REACH, which regulates industrial chemicals, substances with ED properties may be included on the Authorisation List. Use of chemicals included on this list requires that industry proves that the use is safe as a prerequisite for authorisation of that use by the authorities. Under REACH, authorities may also require any information necessary when they have concerns regarding the safety of a substance e.g. due to suspicion of ED related effects. This is done by inclusion of the

substance on the Community Rolling Action Plan List (CoRAP List) with specified information requests. The first CoRAP List will be established in 2012, so at this point in time it is not known how many chemicals may be included.

2.3 Impact assessment of surface waters and effluents

33. Fish toxicity data may be used (together with other toxicity data) to set substance-specific 'benchmark' concentrations (known as environmental quality standards, water quality criteria, or similar) for comparison with monitoring data to assess and manage the quality of water bodies. The principles of their derivation are similar to those for establishing PNECs for risk assessment purposes, but their numerical value may differ due to differences in objectives or practical measurement issues e.g. compliance checks in monitoring programmes.

34. *Ex situ* ecotoxicity assays, some of which involve fish, are used by environmental managers in many countries for assessing the toxicity of both effluents and surface waters (e.g. den Besten and Munawar, 2005). This type of work is sometimes undertaken as part of a discharge consent for complex effluents, and sometimes to assist in identifying toxic substances responsible for aquatic wildlife kills or other impacts (e.g. in Toxicity Identification Evaluation (TIE) schemes, employing repeated sample fractionation followed by bioassay to identify the toxic fractions (e.g. Norberg-King et al. 2005). Much of this work is conducted with organisms which are compact, easy to deploy and give a rapid response, features which often (but not always) rule out the use of fish. Furthermore, such bioassays do not usually employ OECD TGs without significant modification. However, a brief description will nevertheless be given of the various types of fish tests which are deployed in impact assessment.

35. For the purposes described above, probably the most frequently used fish bioassay is some variant of OECD TG 203, i.e. an acute lethal test. One example of this approach is an assessment of the toxicity of mine drainage which used fathead minnow (*Pimephales promelas)* mortality to investigate the effectiveness of remediation of mine effluents (Deanovic et al. 1999). Similar methods have been described by Farré and Barcelo (2003). In Europe, effluents are tested using zebrafish embryos (eggs), and guidelines are available on a national (e.g. Germany, DIN 2001) and international level (ISO 2007). 'Chronic' fish tests with apical endpoints (e.g. growth and development), which may be some variant of OECD TGs 210, 212 or 215, are also sometimes used in assessments of effluent or surface water toxicity, but their relatively high cost and logistic difficulty limits their application considerably.

36. In contrast, fish bioassays employing a variety of biomarkers rather than apical endpoints are widely used for water quality assessment. Many of these are outside the scope of current OECD TGs, but some have close similarities. One example of this is a programme to investigate the estrogenicity of effluents in North America (Huggett et al. 2003). This study deployed inter alia a 7d test with Japanese medaka (*Oryzias latipes*), in which estrogenicity was detected by measuring vitellogenin (VTG) induction in males. The exposure time was shorter than the 21d duration of the OECD fish screening assay (OECD TG 230), but the principles behind the two approaches are identical. Similar work by Barber et al. (2007) exposed fathead minnow (*Pimephales promelas*) to treated estrogenic sewage effluent and not only measured VTG induction, but also a range of endpoints relevant for reproduction such as secondary sexual characteristics, gonado-somatic index and reproductive competency, in a test similar to the fish short-term reproduction assay (OECD TG 229).

37. Many other examples could be given for fish being used in water quality assessment, but, as stated above, they rarely follow OECD guidelines precisely, and so fall outside the scope of this document.

2.4 Summary of hazard/risk assessment requirements and recommendations

38. This brief survey of the fish testing requirements of various chemical regulatory jurisdictions shows that most of the OECD fish-based TGs are required and/or employed under certain circumstances in the various OECD countries. However, until now, the most frequently used OECD TGs are TGs 203, 210 and 305. Given the potential need to assess substances for endocrine active properties, it is expected that OECD TGs 229 and 230 will also come into wide use before long. The fish sexual development test (TG 234) which provides data on both endocrine activity and adverse effects (altered sex ratio) may, in addition, be performed for testing some compounds with suspected endocrine activity. There is clearly a demand for fish life-cycle testing, so it will be desirable for the Japanese medaka multi-generation test (MMGT) to be developed into an OECD TG. However, under many circumstances, a partial- or 1-generation fish life-cycle test (i.e. FFLCT) is likely to be sufficient to satisfy regulatory requirements; so the development and validation of such guidelines are also being considered.

39. It should be noted that currently there appears to be limited demand for OECD TG 204, TG 212, and TG 215; thus, the retention of these guidelines should be evaluated. Possible courses of action could include dropping them entirely, or possibly including them in other guidelines as 'special adaptations, or modifications'.

2.5 References

Barber, L.B., Lee, K.E., Swackhamer, D.L., Schoenfuss, H.L. (2007) Reprodutive responses of male fathead minnows exposed to wastewater treatment plant effluent, effluent treated with XAD8 resin, and an environmentally relevant mixture of alkylphenol compounds. Aquat. Toxicol. 82: 36-46.

Benoit, D.A. (1982) User's guide for conducting life-cycle chronic toxicity tests with fathead minnows (*Pimephales promelas*). Environ. Res. Lab. Duluth, MN. EPA 600/8-81-011.

Deanovic, L., Connor, V.M., Knight, A.W., Maier, K.J. (1999) The use of bioassays and toxicity identification evaluation procedures to assess recovery and effectiveness of remedial activities in a mine drainage-impacted stream system. Arch. Environ. Contam. Toxicol. 36: 21-27.

Den Besten, P.J., Munawar, M. (2005) Ecotoxicological testing of marine and freshwater ecosystems: emerging techniques, trends and strategies. Taylor and Francis, Boca Raton, 271 pp.

DIN (2001). German standard methods for the examination of water, waste water and sludge - Subanimal testing (group T) - Part 6: Toxicity to fish. Determination of the non-acute poisonous effect of waste water to fish eggs by dilution limits (T 6). DIN 38415-6. German Standardization Organization.

EMEA (2004) Guideline on the environmental impact assessment for veterinary medicinal products, Phase II. EMEA/CVMP/VICH/790/03, European Agency for the Evaluation of Medicinal Products, London, 39 pp.

EMEA (2006) Guideline on the environmental risk assessment of medicinal products for human use. EMEA/CPMP/SWP/4447/00, European Agency for the Evaluation of Medicinal Products, London. 12 pp.

European Chemicals Agency (2008) Guidance on information requirements and chemical safety assessment. Chapter R. 7b: endpoint-specific guidance. Guidance for the implementation of REACH. European Chemicals Agency, Helsinki. 234 pp.

European Chemicals Agency (2009) Guidance on the Application of the CLP Criteria. Guidance to Regulation (EC) No. 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures. European Chemicals Agency, Helsinki, 528 pp.

EU (1998), Biocidal Products Directive 98/8/EC.

EU (2002) Guidance document on aquatic ecotoxicology in the context of the Directive 91/414/EEC. Sanco/3268/2001 rev. 4 (final), 62 pp.

EU (2006), Regulation (EC) No. 1907/2006 concerning the registration, evaluation, authorization and restriction of chemicals (REACH), Official Journal of the European Union, L 396, Brussels.

EU (2009), Regulation (EC) No. 1107/2009 of the European Parliament and the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC, Official Journal of the European Union, L 309/1, 1-50, Brussels.

EU (2010), Guidance document on aquatic ecotoxicology, Sanco/11843/2010 (draft), rev. July 2010

Farré, M., Barcelo, D. (2003) Toxicity testing of wastewater and sewage sludge by biosensors, bioassays and chemical analysis. Trends Anal.Chem. 22: 299-310.

FDA-CDER (1998).Guidance for industry – environmental assessment of human drugs and biologics applications. Revision 1. FDA Center for Drug Evaluation and Research. Rockville, Arkansas.

Huggett, D.B. Foran, C.M., Brooks, B.W., Weston, J., Peterson, B., Marsh, K.E., La Point, T.W., Schlenk, D. (2003). Comparison of *in vitro* and *in vivo* bioassays for estrogenicity in effluent from North American municipal wastewater facilities. Toxicol. Sci. 72: 77-83.

ISO (2007) Water quality – Determination of the acute toxicity of water to zebrafish eggs (*Danio rerio*). ISO 15088:2007, Geneva, Switzerland.

Länge, R., Hutchinson, T.H., Croudace, C.P., Siegmund, F., Schweinfurth, H., Hampe, P., Panter,G.H. andSumpter, J.P. (2001). Effects of the synthetic estrogen 17-alpha-ethinylstradiol on the life-cycle of the fathead minnow (*Pimephales promelas*). Environ. Toxicol. Chem. 20: 1216-1227.

Maltby, L., Blake, N., Brock, T.C.M., Van den Brink, P.J. (2005) Insecticide species sensitivity distributions: Importance of test species selection and relevance to aquatic ecosystems. Environ. Toxicol. Chem. 24, 379-388.

McKim, J. M. (1977).Evaluation of tests with early life stages of fish for predicting long-term toxicity. J. Fish. Res. Board Can. 34: 1148-1154.

Norberg-King, T.J., Ausley, L.W., Burton, D.T., Goodfellow, W.L., Miller, J.L., Waller, W.T. (2005) Toxicity reduction and toxicity identification evaluations for effluents, ambient waters, and other aqueous media. SETAC Press, Pensacola, 455 pp.

NICNAS (1990) Australian Industrial Chemicals (Notification and Assessment) Act 1989. http://www.comlaw.gov.au/comlaw/Legislation/ActCompilation1.nsf/0/734FC42762DA0043CA2573550020B79A?OpenDocument

OECD (2002) Fish two-generation test guideline. Draft proposal for a new guideline. Organisation for Economic Cooperation and Development, Paris, 18 pp.

OECD (2006) Fish embryo toxicity (FET) test. Draft proposal for a new guideline. Organisation for Economic Cooperation and Development, Paris, 11 pp.

OECD (2007) Phase 1 of the validation of the fish sexual development test for the detection of endocrine active substances. Organisation for Economic Cooperation and Development, Paris, 67 pp.

OECD (2011). Guidance document on the androgenised female stickleback screen. OECD Series on Testing and Assessment no. 148, ENV/JM/MONO(2011)29, Organisation for Economic Cooperation and Development, Paris, 47 pp.

Oris, J. T., Belanger, S. E. and Bailer A. J. (2012) Baseline characteristics and statistical implications for the OECD 210 Fish Early Life Stage Chronic Toxicity Test. Environ Toxicol Chem. 31 (2): 370-6, doi: 10.1002/etc.747.

US EPA (1972) Federal Insecticide, Fungicide and Rodenticide Act (FIFRA). United States Environmental Protection Agency, Washington DC, http://www.epa.gov/compliance/civil/fifra/fifraenfstatreq.html.

US EPA (1976) Toxic Substances Control Act (TSCA). 40 C.F.R. §§700-799, 15 U.S.C §2601 et seq. United States Environmental Protection Agency, Washington, DC, http://frwebgate.access.gpo.gov/cgi-bin/usc.cgi?ACTION=BROWSE&TITLE=15USCC53

VICH (2003). Guideline on environmental impact assessment for veterinary medicinal products Phase II. International Cooperation on Harmonization of Technical Requirements for Registration of Veterinary Medicinal Products (VICH). CVMP/VICH/790/03-final. European Medicines Agency, London.

Wenzel, A, Schäfers, C., Vollmer, G., Michna, H., Diel, P. (2001) Research efforts towards the development and validation of a test method for the identification of endocrine disrupting chemicals. Report for EU DG 24, B6-7920/98/000015.

3. STATISTICAL CONSIDERATIONS

3.1 Outline

40. The statistical methods used to analyse results of regulatory ecotoxicology studies must be consistent with regulatory frameworks, they must be statistically robust to maintain an acceptable power of the assay and maximize efficiency in terms of animal use, time and costs. Different national and regional risk assessment schemes have often been developed to balance these factors in a variety of ways. For example, in Europe, the environmental hazard assessments of industrial chemicals or biocides focus on the calculation of the predicted no effect concentration (PNEC) values. These are typically based on either acute effects concentration (EC_x) type studies or chronic no observed effect concentration (NOEC) type studies, using assessment factors as appropriate (ECHA 2008). For pesticides, in Europe either acute EC_x type studies or chronic NOEC studies are used to derive toxicity exposure ratios that are employed for risk assessment (Directive 91/414; EU 1991) [New regulation 1107/2009 (EU 2009) will apply from 14 June 2011]. In the United States, the regulatory terminology implies the use of chronic NOEC data as a basis for the calculation of hazard quotients used in the risk assessment of pesticides (US EPA 2004). In contrast, chronic exposure in aquatic algae (e.g. OECD TG 201) studies are typically analysed by calculation of an EC_{10} and EC_{20}, which are often used alongside aquatic fish studies in the assessment of risk. Also, some probabilistic risk assessment schemes might require information on the slope and confidence limits of the dose-response curve.

41. Regulatory needs, test designs and statistical methods cannot be considered independently, and the impact of change in one of these factors on the others must be considered carefully. For example, in Chapter 4 it is stated that for endocrine screens (OECD TG 229 and TG 230) "the definitive test exposes fish to a suitable range of concentrations maximizing the likelihood of observing the effect. The important distinction being that achieving a NOEC is not the purpose of the test". This statement accurately reflects the original basis for the design of the test. Yet, as the basis for the original test design fades in memory, there appears to be a tendency to expect the calculation of both EC_x and NOEC values based on the results of the screens. It cannot be assumed that the design of this test can support adequate estimates of EC_{10} and EC_{20} values without adequate studies of the accuracy and precision of estimates. Alternative testing methods may have new or novel strengths/limitations in terms of statistical power compared to standard guidelines (e.g. testing based on Threshold Approach, OECD GD 126, OECD 2010).

42. Statistical methods must be capable of detecting, or modelling, the smallest effects that are biologically meaningful (for discussion, see below). A key issue in the interpretation of fish toxicity tests, as they grow in complexity, is to distinguish between biologically important effects caused by the test chemical *versus* statistically detectable differences. This aspect of ecotoxicology test guideline data interpretation is identical to the principles developed for many mammalian test guidelines in recent years (Länge et al. 2002, Williams et al. 2007). Against this background, a key element of this chapter is to illustrate key principles of requirements for optimal test design and data interpretation (e.g. importance of historical control values in designing a test for endpoints of interest, adequate replication, etc.) that can be used as required for different fish test guidelines.

43. Toxicological endpoints should not be interpreted in isolation from other information relevant to the test. For example, it is usually assumed that the responses will follow an underlying monotone concentration response pattern (i.e. there is a general tendency for the effect to increase as concentration increases) in the absence of compelling

evidence to assume otherwise. Use of the knowledge that responses are likely to follow such a pattern can lead to better statistical tests and allow variations not related to treatment to be identified. For example, this assumption makes the Jonckheere-Terpstra trend test a powerful tool for the calculation of NOEC values and forms the entire basis for calculation of EC_x values.

44. There is some controversy on the question of whether hypothesis testing (NOEC/LOEC [no/lowest observed effect concentration]) or regression (EC_x) is the better way to evaluate toxicity data (e.g. Chapman et al. 1996, Dhaliwal et al. 1997). It is not the intention to replay that debate here. The intention of this chapter is to indicate how best to do each type of analysis and to indicate the types of data and experimental designs under which each type of analysis can be done with reasonable expectation of useful results. Therefore, requirements for the different approaches are considered.

45. The OECD guidance document no. 54 (OECD 2006) describes current approaches to the statistical analysis of ecotoxicity data and should be consulted. However, the recent development of new fish test guidelines (e.g. Fish Sexual Development Test, TG 234) has raised additional specific issues worthy of discussion in addition to some general considerations.

3.2 Biological *versus* statistical significance

46. The question of what magnitude of effect is biologically important to detect or what effects concentration (EC_x) to determine is not a statistical issue. This issue is not unique to fish tests, but is valid for other ecotoxicity test species such as *Daphnia* and algae. Scientific judgement, grounded in repeated observation of the same response in the same species under the same conditions, is required to specify this (i.e. the understanding of historical control data). Statistics provides a means to determine the magnitude of effect that a given experimental design can quantify. To put this another way, once an effect size of biological importance has been determined, it is possible to design an experiment that has a high likelihood of producing the desired information (i.e. whether an effect of the indicated size occurs at some test concentration or what concentration produces the specified effect).

47. The relationship between biological significance and statistical significance can be understood in terms of the magnitude of effect that can be detected statistically. For a continuous response, such as growth or fecundity, this in turn depends on the relative magnitude of the between-replicate and within-replicate variances. The standard error of the sample control mean response is given by:

$$SE = SQRT[Var(Rep)/r+Var(ERR)/rn] = \sigma \cdot SQRT[R/r+1/rn],$$

where σ is the within-replicate or error standard deviation, R is the ratio of the between-replicate variance to the within-replicate variance, r is the number of replicates in the control group and n is the number of fish per replicate, Var(Rep) is the between-replicate variance and Var(ERR) is the within-replicate or fish-to-fish variance.

48. The 95% confidence interval for the mean is, approximately, Mean ± $2 \cdot SE$[2]. It is often convenient to express 2*SE as a percent of the mean, so the 95 % confidence interval for the mean can be expressed as Mean ± P %, where $P = 200 \cdot SE/Mean$. The true mean is

[2] Note that SE is the standard error of the mean, calculated by dividing the sample standard deviation by the square root of the number of observations. It assumes that the data are normally distributed.

statistically indistinguishable from any value in this confidence interval for the sample mean.

49. This means that the smallest treatment effect that can be distinguished statistically is P % of the control mean. This holds for both the NOEC and EC_x approaches.

50. It is then incumbent on the study director to determine the magnitude of effect, Q, which is judged biologically important. For a given experimental design and endpoint, Q is compared to P. If Q > P, then the experiment is suited for the purpose, otherwise not.

51. For example, in a recent fish full life-cycle (FFLC) test, the control mean for percent male offspring was 69 %, with standard error of the mean = 19 %. Thus, the smallest effect that can be found statistically significant is 19 % and EC_x for x < 19 cannot be reliably estimated. Another way of considering this is to observe that the lower bound of the 95 % confidence interval for EC_{10} and EC_{20} is 0. Also, the NOEC is a concentration at which a > 19 % effect was observed. (NOTE: The confidence interval for the difference between the control mean and a treatment mean is actually greater than 19% by a factor of √2). Vitellogenin (VTG) can be another highly variable response (Hutchinson et al., 2006) and only large effects, around 40 %, can be expected to be statistically significant in a practical experiment. Equivalently, EC_{40} might reasonably be estimated, but not EC_{25}.

52. At the other extreme, the standard error of a growth measurement for *Daphnia* is often 2 - 3 % of the control mean, so very small effects can be found statistically significant. For this response, it is quite feasible to estimate EC_5. It is a matter of scientific (non-statistical) judgment whether such small effects are biologically meaningful. Similar findings hold for avian eggshell thickness measurements.

53. From the formula for standard error (SE), it will be evident that there is a trade-off between the number of fish per replicate and the number of replicates per control or test concentration. For example, from the formula, it is evident that if the number of replicates is doubled and the number of fish per replicate is reduced by 50 %, then the second term in brackets is unchanged, but the first term is reduced by 50 %. This might indicate a preference for more replicates of fewer subjects per replicate. However, if Var(REP) is already relatively small, not much is gained by such an approach. Instead, if we cut the number of replicates by 50 %, but increase the number of subjects per replicate by a factor of 4, then the first term remains small but the second term is reduced by 50 %. Thus, whether it is better to have a few replicates with many subjects in each, or many replicates with few subjects in each, depends on the relative magnitude of the two variances. A general rule is that if the ratio R of variances exceeds 0.5, then more emphasis is given to the number of replicates, otherwise, more emphasis is given to the number of subjects within each replicate. For example, shoot heights of some emergent crops (e.g. oat, tomato, rape) tend to have variance ratios exceeding 0.5 (John Green, pers. comm.).

54. As further illustration, recent experiments conducted for the OECD found VTG measurements to be very highly variable and the within-replicate variance ranged between 2 and 10 times that of the between-replicate variance. Thus, good experimental design would call for relatively few replicates with numerous fish in each. In the case of Japanese medaka (*Oryzias latipes*), it was found that a control and three test concentrations with two replicates per control and test concentration and five fish per replicate were adequate to give 80% power to detect a 60 % effect. For fathead minnow (*Pimephales promelas*), four replicates of four fish each in the control and each test concentration were required to give 80 % power to detect a 94 % effect. While test effect sizes may seem large and the number of fish small, there were constraints on the number of replicates and fish that could be

accommodated from practical considerations. Furthermore, such size effects were observed in high test concentrations in these experiments during validation.

55. A complication is when there are multiple endpoints to be analysed in a single experiment. If the experimental design is optimal for one response, it may be sub-optimal for another response. This may mean that only very large effects can be estimated or detected statistically for one endpoint and perhaps very small, biologically unimportant effects will be found statistically significant for another response. It is important to understand this in interpreting the data. A biologically important effect may be missed in the first instance, which should not be interpreted to mean the chemical in question has no effect on that response; while a sound study might be rejected because of a tiny effect found statistically significant. Any statistical result should be interpreted in light of the biologically important effects determined before the experiment was conducted.

56. To design an experiment with high likelihood of detecting a P % effect (or estimating a meaningful EC_P), it suffices to find r and n so that P = 200·SE/Mean. If enough data are available for a particular test guideline, it is a simple matter to construct a table showing 200·SE/Mean for various values of r and n and seeking the most practical combination to decide on a suitable test design based on historical control estimates of the two variances.

3.3 NOEC/LOEC

57. For the purpose of determining an NOEC or an LOEC, it is important to design the experiment so as to be able to have a reasonable chance of finding a biologically relevant effect statistically significant and minimize the chance of finding biologically irrelevant effects statistically significant. These two objectives are somewhat incompatible, and judgement will be useful in reaching appropriate regulatory conclusions. A fish study will have a water control group (dilution water control), and if a solvent is used, a solvent control, and at least one test concentration. Unless the design is for a limit test, there will usually be three or more test concentrations approximately equally spaced on a log scale. With very few exceptions, the control(s) and test concentrations should be replicated. Replicate here refers to the test vessel, not to individual animals unless they are housed individually in a test vessel. The trade-off between number of fish per replicate and number of replicates per test concentration and control will vary according to the response and species, and a power calculation may be needed to determine the best design. Since multiple responses are usually tested from the same experiment, it will often not be possible to design an experiment that is optimal for all responses. Judgement is needed to decide on the most important response(s) and experiments should be designed to provide adequate power (75 - 80 %) to detect biologically relevant changes in those responses. Power simply refers to the probability of finding statistically significantly an effect of a given true magnitude, taking into account variability in the response of interest and variability arising from sampling. It is also important to quantify the size effect likely to be found significant in all other responses. This may indicate that there is a need to rethink the objectives of the experiment. There should be no surprises at the end of the study about what can be analysed, by what test, and with what ability to detect effects.

58. Most responses are analysed using 2-sided tests, unless there is a clear biological reason to expect or be concerned only with changes in one direction (e.g. an increase). Furthermore, for most responses, there is an expectation that the concentration-response is approximately monotone. The effect may be measured as an increase or a decrease in some measurement (e.g., weight might decrease, mortality might increase). That being true, a test that is designed for a monotone trend is more powerful than one that simply compares each

test group to the control independent of effects in other test groups. Thus, there is a preference for a step-down Williams or Jonckheere-Terpstra test over a Dunnett, Dunn, Mann-Whitney or other pairwise test, provided the data are consistent with a monotone concentration-response. All tests referenced in this chapter are discussed in detail in OECD (2006).

59. All statistical procedures are based on some data requirements. In addition to the monotonicity requirement for Williams and Jonckheere-Terpstra for continuous responses and Cochran-Armitage for quantal responses, there are additional requirements. Note that the Jonckheere-Terpstra test is non-parametric. The Williams and Dunnett tests require normally distributed data with homogeneous variances. While these tests are robust against mild violations of these requirements, they are not impervious and some checking of these requirements is appropriate. A visual check from a scatter plot may be sufficient to assess monotonicity and variance homogeneity, and even normality. There are also formal tests for all three and OECD (2006) provides details.

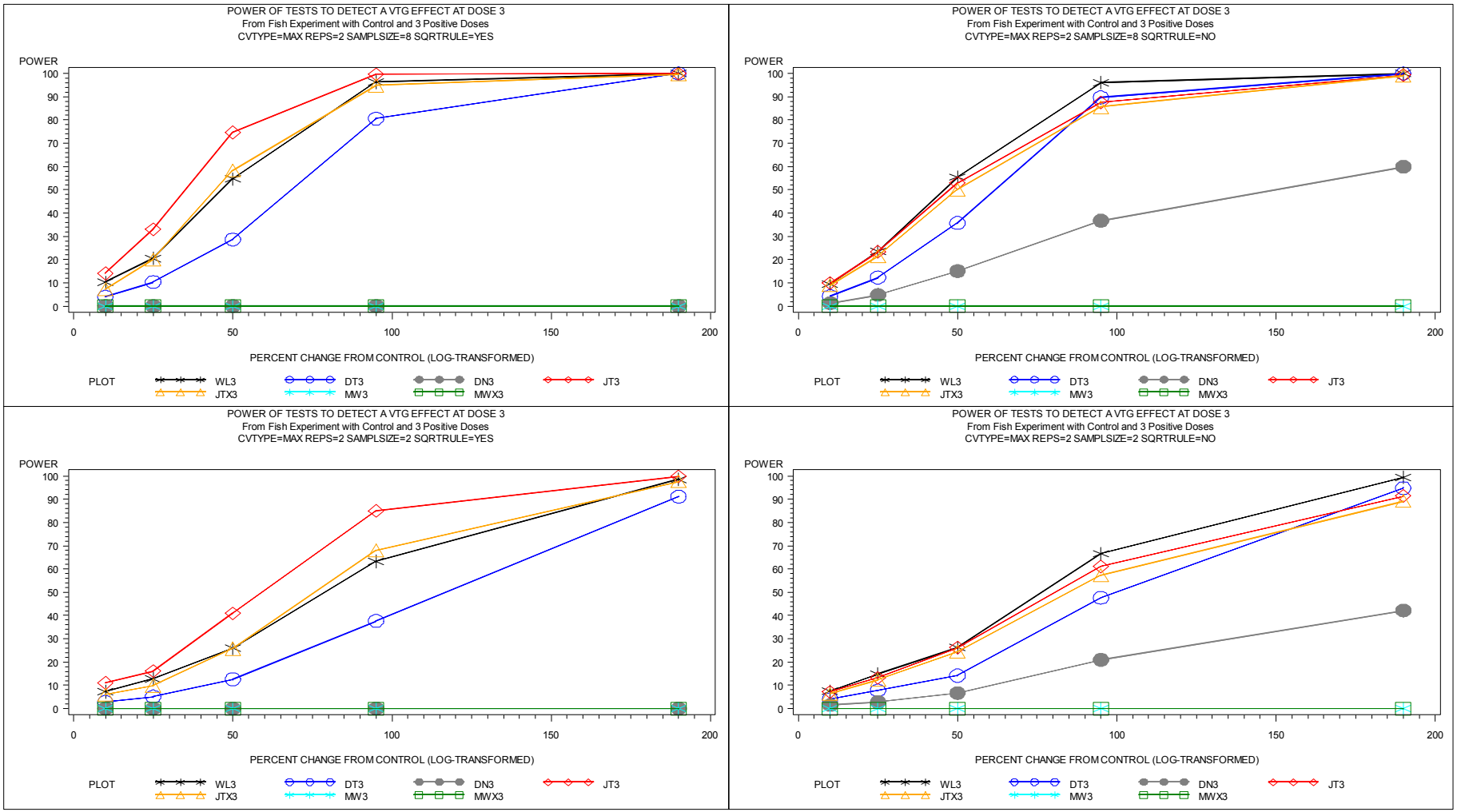
60. Where normality or variance homogeneity are violated, a transformation of the data to achieve these requirements can be sought or non-parametric methods employed, which have fewer requirements or are much less sensitive to violations. Be mindful that different agencies may have different jurisdictions on how, or whether, data will be transformed. Contrary to widely held opinions, non-parametric tests are not always inferior to parametric tests. For example, the power properties of the step-down Jonckheere-Terpstra test are very similar to those of the step-down Williams test, when the data are normally distributed with homogeneous variances, and are superior to Williams, when those conditions are violated. On the other hand, for datasets with few replicates, the power properties of the Mann-Whitney and Dunn tests are worse, sometimes much worse, that those of Dunnett's test. Fig. 3.1 indicates a typical comparison of these tests.

61. Fig. 1 shows the power of seven tests for an experiment with three positive test concentrations and a single control. The horizontal axis shows the percent change from the control, and the vertical axis shows the probability that size effect will be found statistically significant. The red curve shown with diamonds is the step-down Jonckheere-Terpstra (JT) test (standard asymptotic version), orange with triangles for the exact permutation version of JT, the black dark curve with asterisks is William's test, blue with circles is for Dunnett's, cyan with asterisks and green with squares are for the standard (asymptotic) and exact permutation versions of Mann-Whitney (also known as the Wilcoxon) test, and grey dots for Dunn's test. In the top row, the design called for 2 replicates of 8 fish in the control and each test concentration, while the bottom row is for two replicates of two fish each. The left column is for a design following the square-root allocation rule (see below), and the right column is for a design with equal replication in control and all treatment groups. On the left, the gray Dunn power curve is hidden by the green and cyan Mann-Whitney curves. The data generated were normally distributed with homogeneous variances and with variability that was observed for VTG in some OECD validation experiments for fish endocrine screening studies. It is clear that the power of the Jonckheere test is greater than that of Williams test on the left, whereas William's test sometimes has slightly greater power on the right and both tests exceed in power that of all the pair-wise tests (Dunnett, Dunn, Mann-Whitney). A striking feature of these results is that Mann-Whitney has zero power to detect effects regardless of magnitude in either design, whereas Dunn's has zero power under the square-root rule and low power under equal allocation. This knowledge is clearly important in deciding on design and test selection.

3.4 EC_x

62. Standard regression models also depend on data meeting certain requirements. Among these requirements, a key prerequisite is that the observations are mutually independent. This requirement is violated, for example, if all responses are divided by the control mean in an attempt to "normalize" the data or reduce variability. While it is not impossible to model correlated responses, specialized models are required to do so. For a continuous response (i.e. proportion of males is analysed as a continuous response), the data are assumed to be normally distributed with homogeneous variances. There are modifications to accommodate heterogeneous variances, such as alternative variance-covariance structures or weighting. It is also possible to accommodate some types of non-normality.

Fig. 3.1: Power of various tests to detect effects (power= (1-β)*100), β= type II error, making a false negative decision.



63. It is recognized that regression is robust against mild deviations from normality and variance homogeneity, but it can be adversely affected by serious violations of these requirements. Thus, some checking of the distribution of responses is in order, either visually from a scatter plot or through formal testing. For quantal data, normality is not required, but typically the data are assumed to follow a binary distribution within a given treatment group. Quantal data should be checked for extra-binomial variance (more variation than can be accounted for by the simple binomial distribution), the quantal analogue of the homogeneous variance requirement for continuous responses. If extra-binomial variance is observed, there are statistical test methods which take this into account (e.g. Rao-Scott, as described in OECD 2006). Finally, attention should be paid to goodness-of-fit of the model to the data. There are several ways to assess goodness-of-fit. Among the simplest are (1) visual comparisons of the responses predicted by the model to the observed responses, and (2), where replicates are available, comparing the residual mean square from the model against the pure error mean square. If these residual mean squares are significantly larger than the pure error mean square, then the model does not fit the data well. With small datasets typical in this field, this may not be a powerful test. (3) Confidence bounds on the model predictions are very important to show whether the model predictions have any meaning. If no confidence bounds can be produced or they are very wide, then predictions from the model are scientifically unreliable. It should also be understood that typical confidence bounds do not capture model uncertainty, which is one reason for conducting other goodness-of-fit assessments such as items (1) and (2). Model confidence bounds are constructed based on the assumption that the model is correct. If no confidence bounds can be computed or they are very wide, then the model is not internally consistent regardless of how well the model appears to fit the data from a visual inspection. It is also possible to compute an R-square value to judge the proportion of the total sum of squares that is accounted for by the model. While R-square is a useful measure for linear models, it is an unreliable guide for the non-linear models which are most often used to model ecotoxicity responses. For comparing two models for the same data, Akaike's AIC criteria can be useful.

64. In more general terms, a search of OECD TGs 204, 210, 212, 215, 229 and 230 was made to determine procedures specified for EC_x calculation. Only OECD TGs 212 and 215 describe EC_x procedures. OECD TG 212 describes a normalization procedure, but does not specify fitting a regression curve to the normalized percentages from which to estimate EC_x. In contrast, OECD TG 215 describes two test designs: one for EC_x and one for NOEC determination. Acknowledging the necessary differences: "that a design which is optimal (makes best use of resources) for use with one method of statistical analysis is not necessarily optimal for another. The recommended design for the estimation of a LOEC/NOEC would not therefore be the same as that recommended for analysis by regression."

65. If EC_x designs are to be more widely used, consideration of the following will be necessary prerequisites:

- General guidance should be given on how the need to estimate EC_x affects the optimum spacing of concentrations, number of treatments, and number of replicates to be used. This guidance will probably suggest test designs quite different for the minimum designs currently described in the OECD TGs optimized for NOEC determination.

- Different endpoints may elicit responses at very different concentration levels. Therefore, a strategy is required to handle this within one test.

    - What constitutes a meaningful EC_x estimate, and what is the implication, if a meaningful estimate cannot be obtained for one or all endpoints?

66. Validation of the Fish Sexual Development Test (TG 234) raised particular concerns about the use of regression analysis for determination of effects on sex ratio. These issues are explored in detail in Appendix 1 (Section 3.9).

3.5 NOEC versus EC_x designs

67. It is stated in OECD (2008) that: "[92.] In summary, ANOVA designs for fish testing appear inferior to regression designs, and the latter are considered showing more promise for fish life-cycle tests given the generally large inherent variability in egg production (fecundity) between individuals, which inevitably reduces the power of the ANOVA approach. Final decisions on which design strategy to use should be made on a case-by-case basis, taking into account factors such as the known variability in reproductive output of the species in question."

68. The example datasets and analyses discussed in Chapter 3.9 (Appendix) should serve as a caution against overstating the advantages of regression over what is referred to as the ANOVA approach. While regression has always been an important tool for statisticians, it is not appropriate for some datasets and can suggest a level of precision that is not supported by the data. Experiments for regression analysis call for different experimental designs than those for which ANOVA methods are intended. Just as ANOVA methods call for designs with adequate power to detect biologically relevant adverse effects, regression methods call for designs that are capable of providing reliable or meaningful estimates of an x % effects concentration and this requires designing around the specific *x* or percent effect to be estimated. Basic requirements include the following: (1) There should be test concentrations on both sides of EC_x. The zero concentration control does not figure into this requirement because it is not involved in the probit calculation of EC_x. The purpose of the control in such experiments is purely to provide evidence that the test fish are in good condition. However, see the following paragraph. (2) If the 95 % confidence interval for the control response is of the form Mean ± P %, then estimates of EC_x are meaningful only for $x > P$. For example, if the control mean is estimated only with 20 % error, then it is meaningless to estimate EC_{10}. (3) If the confidence interval for EC_x is very wide, perhaps spanning several test concentrations, then there can be little or no confidence in the EC_x estimate. These basic requirements are important to keep in mind, because once a regression model is fit to the data, it is a simple mathematical exercise to use the resulting equation to estimate EC_x for any percent x, and yet not all such values of x lead to plausible or meaningful estimates. A mathematical equation is not a substitute for valid interpretation of data. This is akin to the requirement in the NOEC approach of adequate power to detect an effect of magnitude deemed biologically important.

69. It can be appropriate to determine both an NOEC (provided there is sufficient power to detect biologically relevant effects) and an EC_x (provided x is beyond the range of control variability). Ideally, x should be between two tested concentrations. However, it is important to recognise extrapolation beyond the range of data adds significant uncertainty and needs to be justified. It is permissible to extrapolate modestly beyond the range of tested concentrations, provided this does not violate restriction (2) in the previous paragraph. Such extrapolation necessarily comes with increased uncertainty and assumes the model fit is valid beyond the range of tested concentrations, something that is untestable from the data. The uncertainty increases the further from the experimental range one extrapolates.

70. Although it is not recommended to combine NOEC and EC_x approaches in the same study, there may be some compelling reasons to do so. For certain existing regulatory frameworks, it might be appropriate to focus on NOEC test designs for fish chronic endpoints (e.g. FIFRA: US EPA 1996). However, for future regulatory frameworks it could be required to have both EC_x and NOEC determinations in fish chronic studies (e.g. draft Sanco 2010 review document). However, the latter

has serious implications for experimental design, time and costs, as well as ethical and statistical interpretations. It might not be practical to design tests with multiple endpoints to determine both the NOEC and EC_x values for the endpoints of interest.

3.6 Alternate designs (*e.g.* square root allocation rule)

71. There are several factors that affect the power of a given test. These are experimental design (e.g. number of replicates per control and treatment group, number of fish per replicate, number of treatment groups), shape of the concentration-response, and inherent variability of the response of interest. One simple, but important decision is whether the control and treatment groups should be equally replicated or whether more replicates should be allocated to the control. The argument for the latter is two-fold: First, it gives a better measure of the undisturbed population against which all treatment groups are compared, and, second, it tends to increase the power of the test, in part by increasing the degrees of freedom for the test statistics.

72. Dunnett (1955) showed that the power of his test is optimized using what is called the square-root allocation rule, which provides a specific formula for the number of replicates in the control and all treatment groups. Details are given in OECD (2006). Further theoretical published work and extensive power simulation studies have shown that this same rule (or a minor modification) also maximizes the power of the Williams and Jonckheere tests and usually increases the power of the Mann-Whitney and Dunn tests.

3.7 Solvent/carrier control

73. One of the issues brought up in the first version of the OECD Fish Sexual Development Test Review, and which is a consideration in all of the test guidelines, is how the two controls (dilution water and solvent controls) should be used when there is a solvent used in the treatment groups. There are advantages to pooling the two controls to test for treatment effects: (1) By doubling the number of control replicates, the power of the tests for treatment effects is increased, achieving at least part of the advantages of the square-root allocation rule described above. (2) All the data are used and the pooled control provides the best estimate of the background population from the experiment. Permissible solvents are those which have been well-established in fish experiments and have been found to have no practical effect on fish at the concentrations used. A preferable alternative to always pooling the controls is to compare them statistically and pool them, if no significant difference is found, and otherwise use only the solvent control to test for treatment effects. The justification for the latter is that solvent is in all the treatment groups at approximately the same concentration as in the solvent control, so that one compares solvent *plus* treatment to solvent, the difference being the treatment effect. This is a plausible hypothesis based on the apparent additivity of effects in most aquatic chemical mixtures that is supported by concentration addition. References on this include Belden et al. (2007), Backhaus et al. (2010), as well as Kortenkamp et al. (2007). The last communication addresses endocrine disrupting chemicals specifically as well as other classes of chemicals.

74. Currently, there is a lack of harmonisation amongst different regulatory authorities on what is the control for statistical analysis (dilution water control or solvent control and if they should be pooled or not). A definitive answer to this question cannot be provided at present, but it has been recommended that a working group should be formed to progress this issue. A topic that might be addressed by such a working group is the reduction in the number of animals that could arise by eliminating one of the controls.

3.8 Power

75. In the design stage, the primary use of power analysis in toxicity studies is to demonstrate adequate power to detect effects that are large enough to be deemed important. If our methods have sufficient power, and we find that, at a given concentration, there is no statistically significant effect, we can have some confidence that there is no effect of concern at that concentration. Failure to achieve adequate power can result in large effects being found to be statistically insignificant. On the other hand, it is also true that a test can be so powerful that it will find statistically significant effects of little importance.

76. Deciding on what effect size is large enough to be important is difficult. In some cases, the effect size may be selected by regulatory agencies or may be specified in guidelines.

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

78. At the design stage, power must, of necessity, be based on historical control data for initial variance estimates. It may also be worthwhile to do a *post-hoc* power analysis to determine whether the actual experiment is consistent with the criteria used at the design stage. If there is significantly higher observed variance (e.g. based on a chi-square or F-test) than that used in planning, then the assumptions made at design time may need to be reassessed. Care must be taken in evaluating *post-hoc* power against design power. Experiment-to-experiment variation is expected, and variance estimates are more variable than means. The power determination based on historical control data for the species and endpoint being studied should be reported.

Power Example

79. Suppose we want to determine the NOEC for mortality in an experiment with rainbow trout (*Oncorhynchus mykiss*), where past experience with this species suggests that background mortality rate at the relevant age and test duration is near zero. We want to be able to detect a 20 % mortality rate and, based on preliminary range-finding experiments, we have decided on an experiment with five test substance concentrations at 50, 100, 200, 400, and 800 ppm, plus a single (non-solvent) control. Furthermore, suppose previous experience suggests that extra-binomial variance and within-tank correlations of responses are unlikely, so a standard Cochran-Armitage test can be done treating all fish within a concentration equally (i.e. ignoring any tank or replicate effect). How many fish per concentration should we plan?

80. First, consider designs with the same number, n, of fish in each concentration as in the control. The power of the Cochran-Armitage test depends on the shape of the concentration-response curve, which we do not know. Powers have been simulated for numerous shapes. Based on an examination of the various power plots, a reasonable choice for design purposes is the linear

concentration-response shape. In addition, the power depends on the threshold of toxicity. For design purposes, we will assume that is zero. The following plots will help (Figs. 3.2 and 3.3).

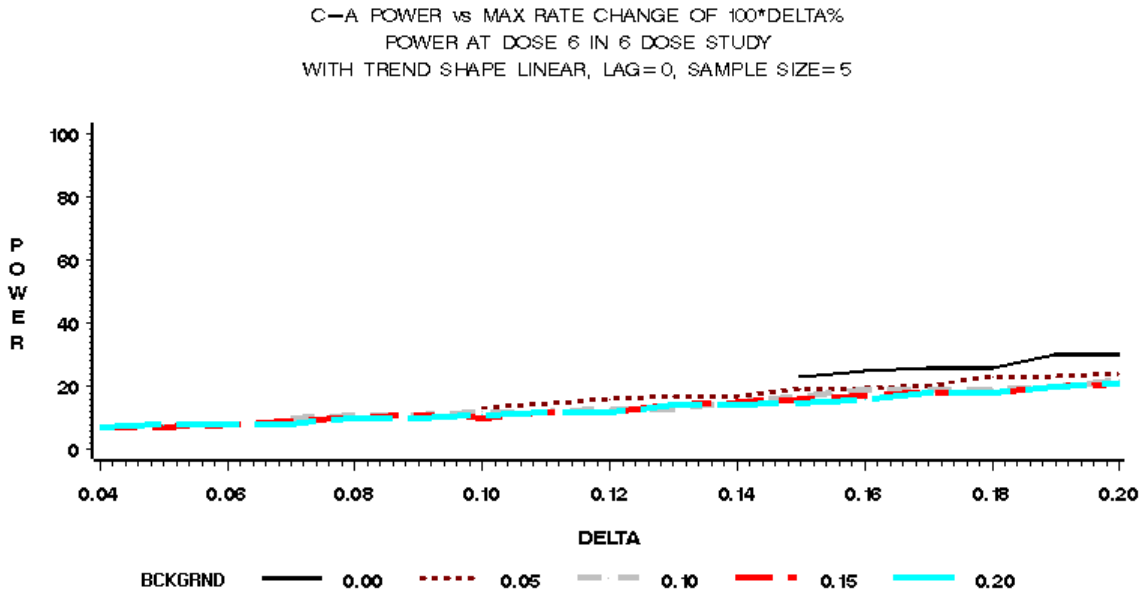


Fig.3.2: Power versus maximum change of 100·Delta % for n = 5 fish per concentration.

81. Fig. 3.2 shows that 5 fish per concentration would give very low power (about 25 %) to detect a 20 % change in the high concentration. There is little point in conducting the experiment for the purpose.

82. Consider a design with 20 fish per concentration: This sample size gives a power of 82 % to detect a 20 % mortality in the 800 ppm concentration (Fig. 3.3). This may well be adequate. What is the power to detect a 20 % mortality rate in lower concentrations? Fortunately, we do not lose much power as we step down. The power to detect a 20 % mortality rate at 400 ppm is 80 %, at 200 ppm it is 78 %, and at 100 ppm it is 76 %. Notice, however, that if the background incidence rate were 10 %, then the power to detect an increase in mortality rate of 20 % drops to around 40 %, which would be inadequate for most purposes.

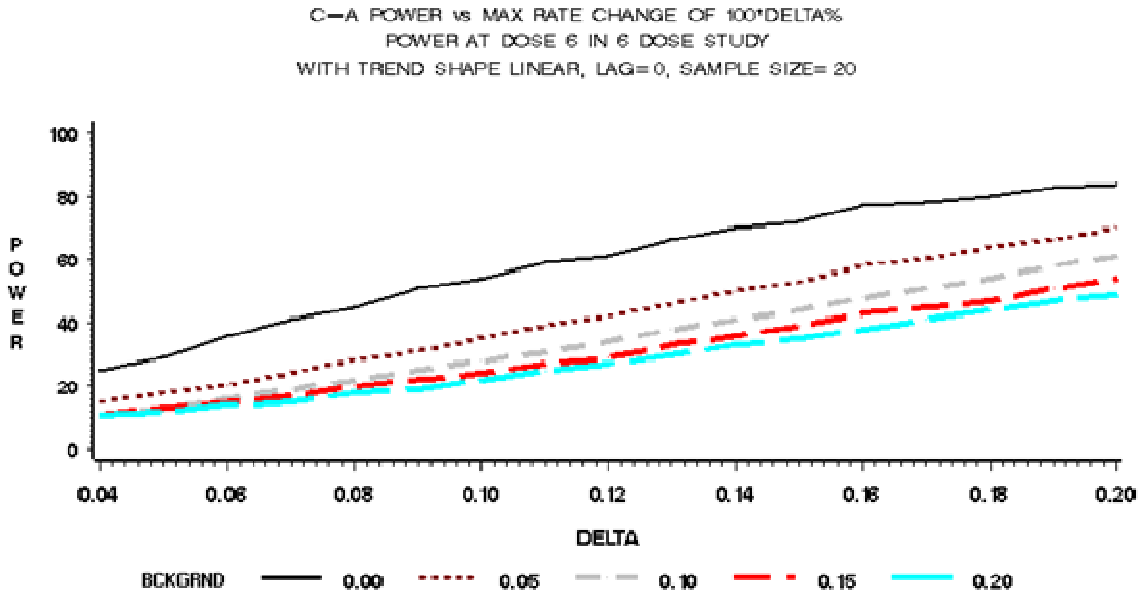


Fig.3.3: Power versus maximum change of 100·Delta % for *n* = 20 fish per concentration.

3.9 Replicates

83. Decisions on the number of fish per tank and number of tanks per group should be based on power calculations using historical control data to estimate the relative magnitude of within- and among-tank variations and correlations. If there is only one tank per test concentration, then there is no way to distinguish housing effects from concentration effects and neither between- or within-group variances or correlations can be estimated, nor is it possible to apply any of the statistical tests described for continuous responses to tank means. Thus, a minimum of two tanks per concentration is recommended; three tanks are much better than two; four tanks are better than three. Some non-parametric tests (e.g. Mann Whitney) require a minimum of four replicates. The improvement in modelling falls off substantially as the number of tanks increases beyond four. (This can be understood on the following grounds: The modelling is improved, if we get better estimates of both among- and within-tank variances. The quality of a variance estimate improves as the number of observations on which it is based increases. Either sample variance will have, at least approximately, a chi-squared distribution. The quality of a variance estimate can be measured by the width of its confidence interval and a look at a chi-squared table will verify the statements made). For further discussion and other tests (parametric and non-parametric), see OECD (2006).

84. The number of tanks per concentration and fish per tank should be chosen to provide adequate power to detect an effect of magnitude judged important to detect. This power determination should be based on historical control data for the species and endpoint being studied.

85. Since the control group is used in every comparison of treatment to control, consider allocating more fish to the control group than to the treatment groups in order to optimize power for a given total number of fish. The optimum allocation depends on the statistical test to be used. A widely used allocation rule was given by Dunnett (1955), which states that for a total of n fish and k treatments to be compared to a common control, if the same number, n, of fish are allocated to every treatment group, then the number, n_0, to allocate to the control to optimize power is determined by the so-called square-root rule. By this rule, the value of n is (the integer part of) the solution of the equation $N = kn + n\sqrt{k}$, and $n_0 = N - kn$. [It is almost equivalent to say $n_0 = n\sqrt{k}$.]. This has been

shown to optimise power for Dunnett's test. It is used, often without formal justification, for other pairwise tests, such as the Mann-Whitney and Fisher exact test. Williams (1972) showed that the square-root rule may be somewhat sub-optimal for his test and optimum power is achieved when $\sqrt{k}$ in the above equation is replaced by something between $1.1\sqrt{k}$ and $1.4\sqrt{k}$.

86. Computer simulations show that for the step-down Jonckheere-Terpstra and Cochran-Armitage tests, power gains of up to 25 % can be realized under the square-root rule compared to results from equal sample sizes.

Power example, continued

87. What if we used the square-root rule in the above power example? Based on the above, we will examine the case where a total of 120 fish are used (20 per concentration and control in the above design). Under the square-root rule, we solve $120 = 5n + n\sqrt{5}$ for n to get n = 16. Then $n_0 = 120 - 5 \cdot 16 = 40$. The following power plot is based on this allocation (Fig. 3.4). Note that the power to detect a 20 % increase in mortality rate in the 800 ppm group is now 92 %. So, with the same number and spacing of concentrations and the same total number of fish, the power to detect a 20 % increase in mortality rate has increased from 82 % to 92 % by using the square-root allocation rule instead of equal sample sizes. An alternative way to use the square-root rule would be to reduce the total number of fish required without loss of power. Indeed, power curves for nominal sample size N = 15 under the square-root rule show the power to detect a 20 % increase in mortality is 86 %. Thus, with a smaller total number of fish allocated optimally, the power to detect a 20 % increase is actually increased. This result underscores the importance of good experimental design and test selection.

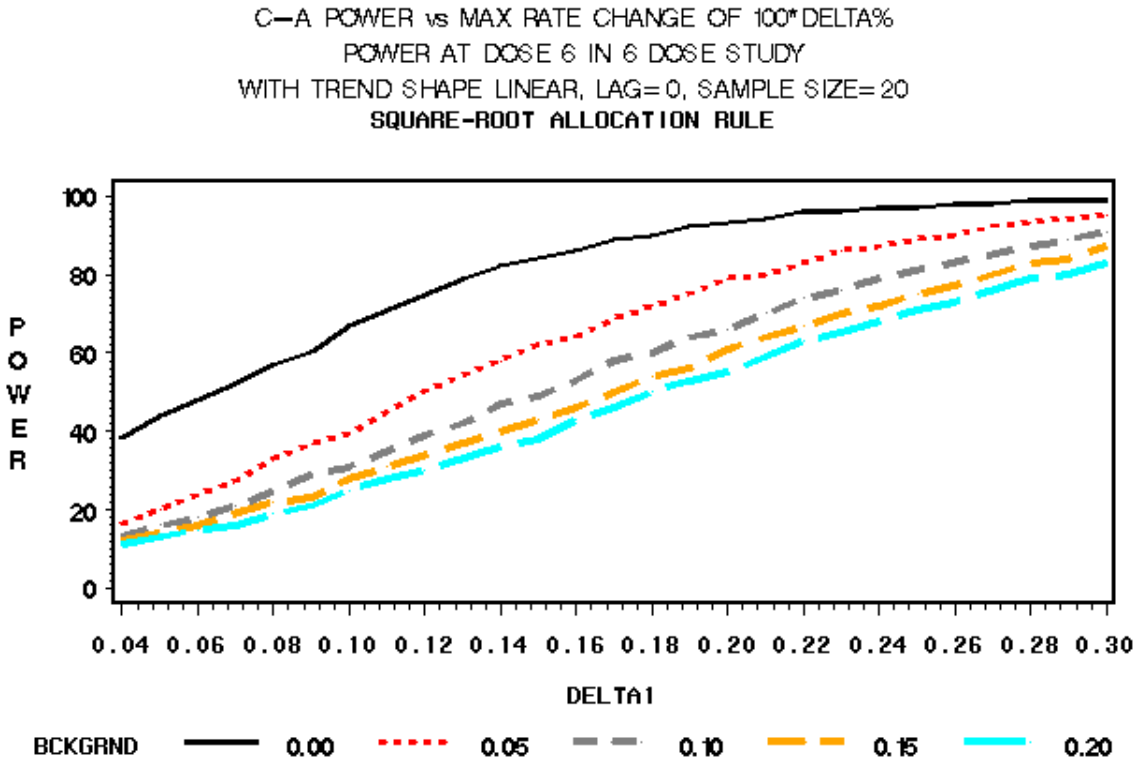


Figure 3.4: For details, see text.

88. In experiments where two controls (dilution water and solvent controls) are used and controls are combined for further testing, a doubling of the control sample size is already achieved. Since experience suggests that most experiments will find no significant difference between the two controls, the optimum strategy for allocating fish is not necessarily immediately clear. This of course would not be a consideration, if a practice of pooling of controls is not followed.

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

90. The square-root allocation rule holds little, if any, advantage for regression analysis. The reason for this is that the curve fitting activity only happens once, with all data. There is no special consideration or use of the controls.

91. The fish toxicity assay most commonly used to estimate chronic effects is the OECD 210 Fish Early Life Stage Test. A systematic analysis of the experimental design and statistical characteristics of the test was undertaken by compiling data compiled from > 100 OECD 210 tests conducted by industry labs (Oris *et al*., 2012). The distribution of responses observed in control treatments was analyzed, with the goal of understanding the implication of this variability on the sensitivity of the OECD 210 TG and providing recommendations on revised experimental design requirements of the test. Studies were constrained to fathead minnow, rainbow trout, and zebrafish. Dichotomous endpoints (hatching success and post-hatch survival) were examined for indications of over-dispersion to evaluate whether significant chamber-to-chamber variability was present. Dichotomous and continuous (length, wet weight, dry weight) measurement endpoints were analyzed to determine minimum sample size requirements to detect differences from control responses with specified power. Results of the analysis indicated that sensitivity of the test could be improved by maximizing the number of replicate chambers per treatment concentration, increasing the acceptable level of control hatching success and larval survival compared to current levels, using wet weight measurements rather than dry weight, and focusing test effort on species that demonstrate less variability in outcome measures. From these analyses evidence was provided to inform the impact of expected levels of variability on the sensitivity of traditional OECD 210 studies and the implications for defining a target for future animal alternative methods for chronic toxicity testing in fish. Power analyses indicated that zebrafish assays had greater statistical power than fathead minnow which were more powerful than rainbow trout (examples for hatching success and post-hatch survival in Fig. 3.5). However, this does not suggest the order of toxicological sensitivity. A separate analysis, in development, addresses this through comparisons (based on the same data set) of NOECs and EC_x determinations for all endpoints. Of these, length and wet weight were more sensitive than other apical endpoints.

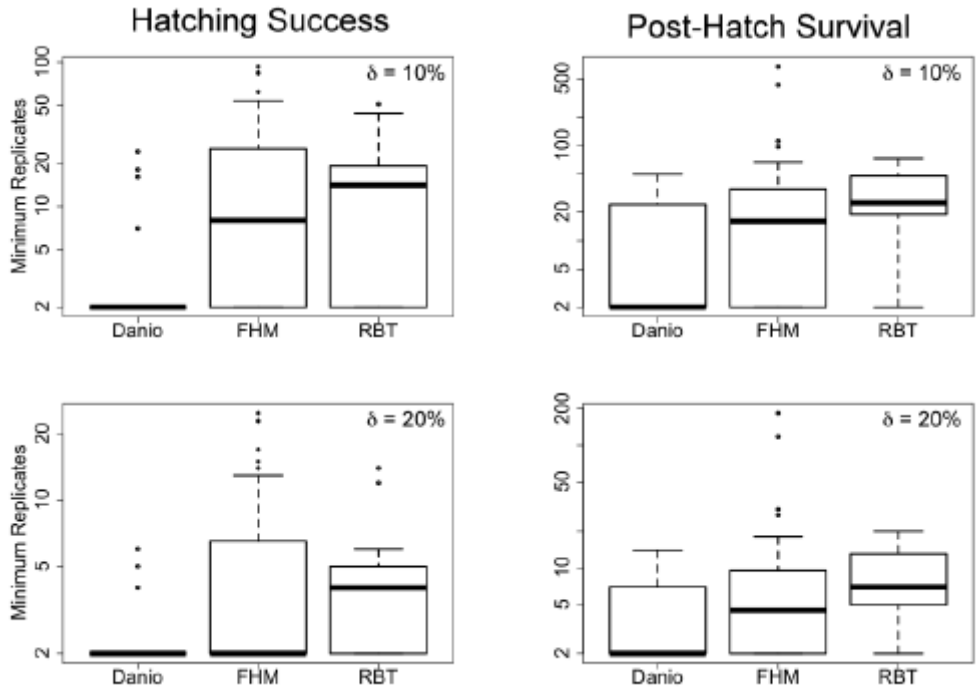


Figure 3.5. Results of power analysis for hatching success and post-hatch survival in the OECD 210 Chronic Fish Toxicity Test. Analyses were conducted on each individual test assuming a minimum control performance of 80% for hatching or survival, with α=0.05 and β = 0.2 (Power = 1- β)=0.8). Plots are shown for the distribution of minimum sample size (i.e., chambers per treatment) required to detect differences of 10% (δ=10%) or 20% (δ=20%) in experimental treatments compared to controls. Boxplots show the median (dark bar), a box bounding the 25th –75th percentile, whiskers spanning 1.5x of the interquartile range, and dots indicating potential outliers in the database (Oris *et al*., 2012).

3.10 Detailed consideration of regression analysis for sex ratio endpoints

3.10.1 Comparison of alternative models

92. When using regression models based on treating the proportion male or female as a continuous response, there will be a need to select one model from among several candidate models fit to the data. There are both formal and informal selection criteria appropriate for the class of models that will be used for sex ratio data. The simplest approach to comparing models should be used in all cases, even when formal tools will also be used. This is visually inspecting the fits to the data. It is important to identify regions of concentrations where each model provides a poor fit. Next, the widths of the confidence bounds about the fitted curves should be compared. Generally, a model that gives narrower confidence bounds is preferred, but this does not outweigh the fit of the model to the data. There are situations where narrow confidence bounds are obtained for a model that clearly does not fit the data. Next, confidence bounds for all estimated parameters should be examined. If the confidence interval for a model parameter contains zero, then the model is suspect, as that parameter is evidently not required. Beyond that, preference is given to the model where the confidence intervals for the parameters are smallest. Where replicate data are available, the residual mean square from the model should be compared to the pure error mean square, which can be obtained from an ANOVA. Finally, Akaiki's or Schwartz's information criterion can be used. The preferred form for Akaiki's information criterion is given by the formula below. A good discussion of AICc is given in Motulsky and Christopoulos (2004).

$$AICc = Ln(RSS/n) + \frac{n+k}{n-k-2},$$

where RSS is the residual sum of squares from the model, n is the total number of observations, and k is the number of parameters estimated for the model.

93. In general, the model with the smaller AICc is preferred. If the values of AICc are close for two models, it is helpful to compute the probability that model A is better than model B using the following formula:

$$\Pr ob = \frac{e^{-D/2}}{1+e^{-D/2}},$$
$$where \quad ,$$
$$D = AICc(B) - AICc(A)$$

94. Here AICc(B) denotes the value of AICc for model B. If the probability is high, then model A is favoured. The following plot of these probabilities may be helpful (Fig. 3.6).

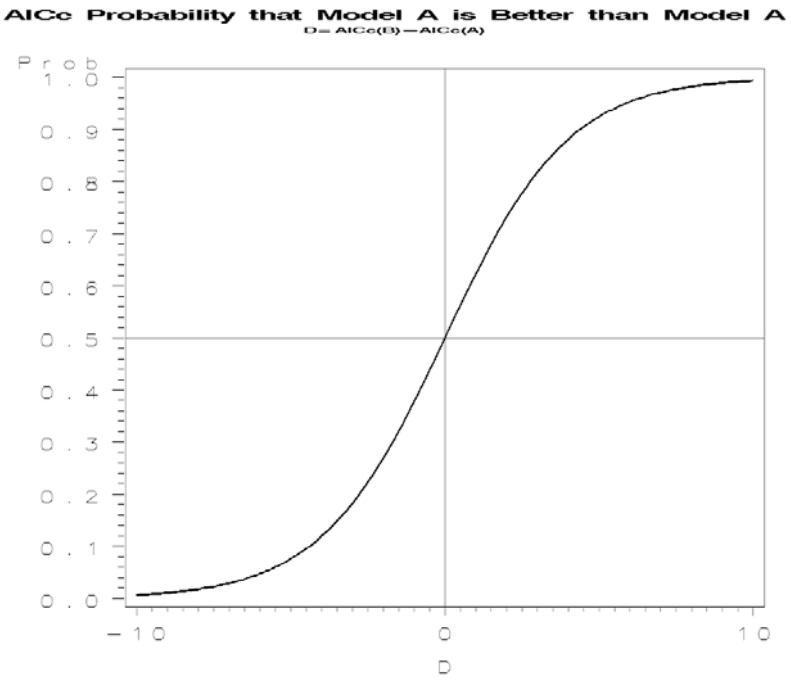


Fig. 3.6: Probability of correct model selection

95. It will be observed that if AICc(B) - AICc(A) ≥ 10, the model A is almost certainly better than model B. This criterion is limited, however, where weighted fits are used, as two models can be compared using this criterion only if they use the same weights. So, in comparing two unweighted model fits (i.e., weight=1), the criterion is sound. For weighted models where the weights depend on the function being estimated, the criterion is not appropriate and comparing an unweighted to a weighted model fit is certainly not appropriate. Some discussion of weights in using AICc is given in http://www.boomer.org/manual/ch05.html a**nd** http://www.micromath.com/products.php?p=scientist&m=statistical_analysis**.**

3.10.2 The meaning of an x % effect

96. The third issue concerns the meaning of an x % effect. For incidence data (such as percent males), there are three distinct concepts that are sometimes confused:

- Absolute risk is when x % of the population is affected.
- Additional risk is when x % above the "background" is affected, so that if the background incidence rate is c %, then the total risk is (x+c) %.
- Relative risk is when x % of the population that would "normally" not be affected is affected, or c % + (1 - c/100) · x %.

97. To illustrate the difference, consider the meaning of EC_{50} when the background incidence is 20 %, i.e., C = 0.2 (background incidence rate).

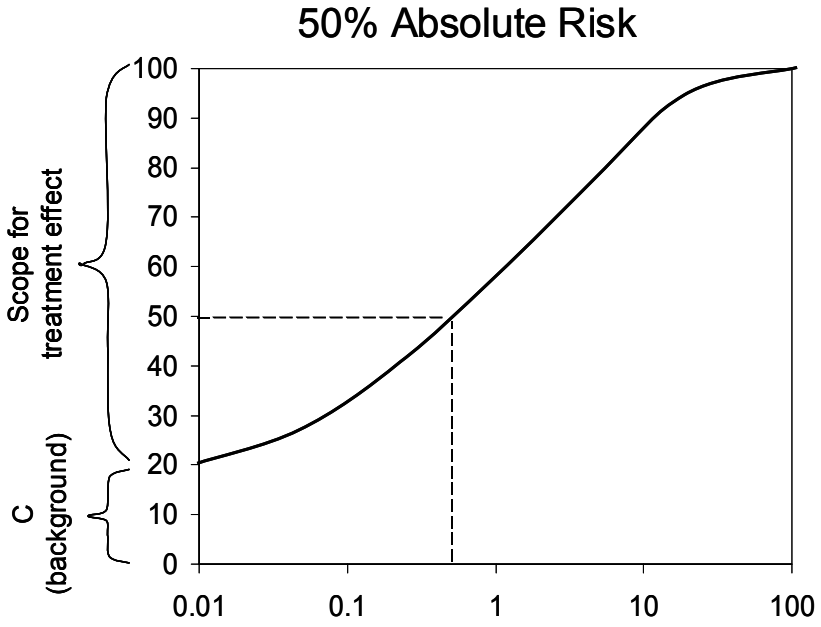


Fig. 3.7: Absolute risk: 50 % of the population is affected (only 30 % above background)

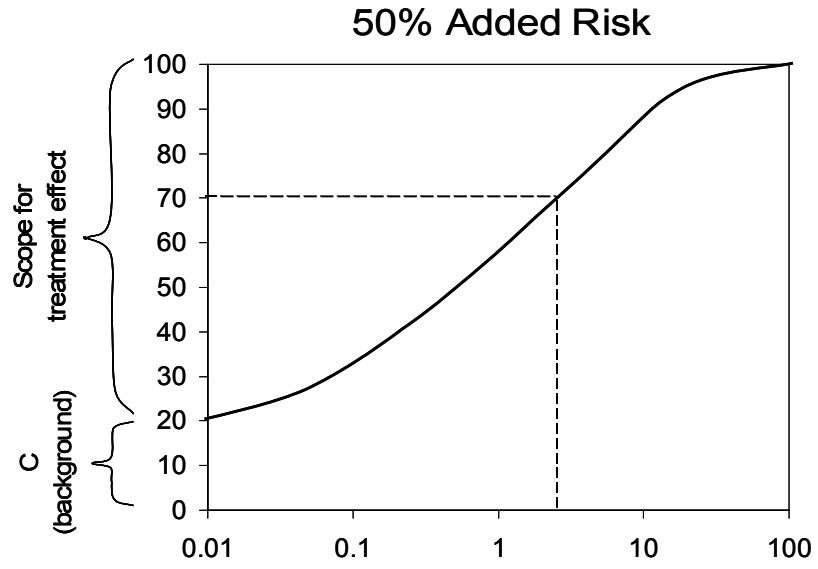


Fig. 3.8: Added risk: 70 % is affected (total = 20 % + 50 %)

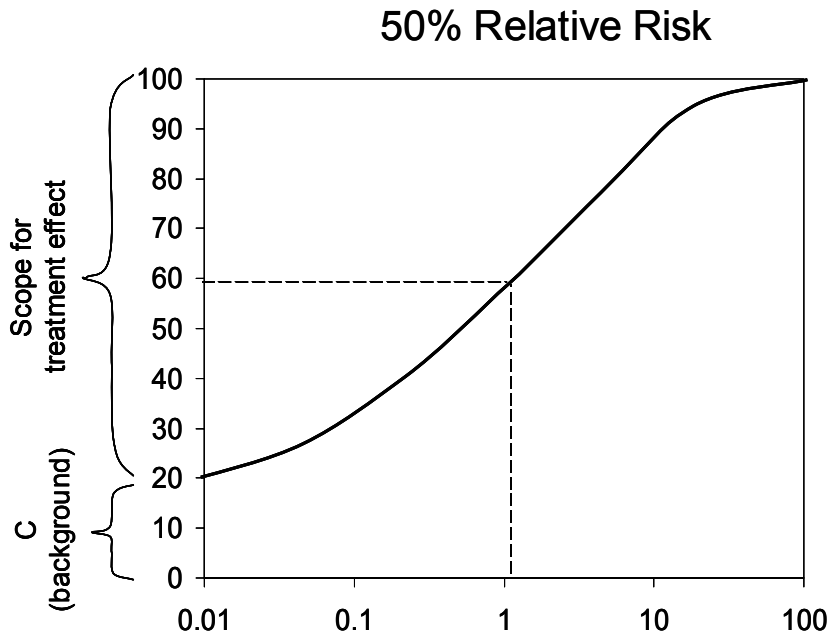


Fig. 3.9: Extra (or relative) risk: 60 % is affected (total = 20 % + 50 · (1 - 0.2) % = 20 % + 40 % = 60 %)

98. Probit analysis of incidence data is based on the concept of relative risk and care must be taken to arrive at the correct EC_x estimate, if there is background incidence. For the sex ratio, if there is no background proportion male (i.e. no males in the control) and there are only males and females, then EC_{50} for males is the same as EC_{50} for females. However, if there is a background incidence of males, the two approaches are not equivalent. This is because it is important in probit analysis to analyse an *increasing* proportion, when there is background incidence. Probit analysis can handle background incidence for an increasing function, but gets thoroughly confused accounting for background in a decreasing function, for what does a "background" incidence rate of 70 % mean when the incidence rate in a treatment group is 40 %?

3.11 Glossary of statistical terms used

ANOVA: Analysis of variance

NOEC – No Observed Effect Level

EC_x – Effective Concentration at which 50% of the effect measured (e.g. mortality) is observed

Almost all of the tests listed below are discussed in depth with references and examples in Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. Organisation for Economic Co-operation and Development. Report no. ENV/JM/MONO(2006)18 Number 54. 1-147. That should serve as the primary reference. The Rao-Scott Cochran-Armitage test is not well covered there and additional references for it are provided below in the description of that test.

Chi-square - Definition and reference to the statistical test

Test for quantal data (i.e., count data, the number of affected organisms compared to the number of tested organisms). The chi-square test is a general test for differences among groups. In its simplest

use, each organism can have two states, affected and not affected. In more complex studies, each organism can have multiple states, such as severity scores in histopathology. A given group (treatment or control) will have some proportion of organisms in each state. The chi-square test is intended to test for whether there are any differences in these proportions from one group to another. No consideration is given to whether a group is a control or treatment, so a significant chi-square test may mean there are two treatment groups that are different when neither is different from the control. In cases where there are multiple possible states, a significant chi-square test can mean that one group is more diverse than another. Conversely, there can be a treatment group mean that is significantly different from the control mean and the chi-square test is not significant. This can happen because the chi-square must "guard" against many comparisons of no interest in toxicology and so can miss an effect that is important in that field. Partly for that reason, a significant chi-square should never be required in order to use one of group comparison tests described below, which are attempts to focus attention on comparisons of interest to toxicology, namely comparisons of treatments to control.

Cochran-Armitage – Definition and reference to the statistical test

Test for quantal data (i.e., count data, the number of affected organisms compared to the number of tested organisms). Quantal (binary) data can be collected and categorized by explanatory factors (such as dosage or treatment level). An analysis of such data usually tries to indicate relationships between the response (binary) variable and factors such as dose level. In such cases, the Pearson Chi-Square (χ^2) test for independence can be used to find if any relationships exist. The Cochran-Armitage test decomposes the Pearson Chi-Square test into a test for linear trend for the dose-response and a measure of lack of monotonicity, $\chi^2_{(k-1)} = \chi^2_{(1)} + \chi^2_{(k-2)}$ where $\chi^2_{(1)}$ is the *1* df calculated Cochran-Armitage linear trend statistic and $\chi^2_{(k-2)}$ is *k-2* df Chi-Square test statistic for lack of monotonicity. While technically, the test is for linear trend, in fact any monotone trend is well-handled by this test. The Cochran-Armitage test can be applied in step-down fashion by first applying it to all the data. If that test is significant at the 0.05 level, then the high test concentration or dose is dropped and the test repeated. This process is continued until the first non-significant test is obtained. The highest dose or concentration remaining at that step is the NOEC. All tests are conducted at the 0.05 level and the over-all false positive rate in this process is 0.05.
The Cochran-Armitage Chi-square can also be expressed as a z-statistic in order to take account of the direction of the trend. The z-statistic is obtained from the formula given by removing the exponent 2 in the numerator and taking the square-root of the denominator. This z-test has a standard normal distribution under the null hypothesis of no trend, and the probability of the z – statistic can be obtained from a table of areas under the standard normal distribution. Only the z-statistic is appropriate for 1-sided tests. Unlike the 1-sided test, the χ^2 version of the Cochran-Armitage test can remain significant in a step-down application even when there is a change in direction of the trend. To avoid this situation when doing a 2-sided test, one applies both 1-sided z-tests with all doses present at the $\alpha/2$ level. At most one of these can be significant. If one is significant, this determines the direction of the trend and all further tests are done with the z-statistic for that same direction at the $\alpha/2$ level.

Dunn (continuous response)– Definition and reference to the statistical test

While Dunn's test is often described in texts as a way of comparing all possible pairs of treatment groups, her original paper (Dunn, 1964) provided a means of estimating any number of general contrasts and adjusting for the number of contrasts estimated. For present purposes, we shall describe only comparisons of treatments to a common control. The procedure is to rank the

combined treatment groups, using the mean rank for tied responses. Compute the mean rank, R_i, for each treatment group. Compute the combined sample size, N, and the individual sample sizes n_i. Finally, compute the variance of $R_i - R_0$ as follows.

$$V_{i0} = \left[\frac{N(N+1)}{12} - \frac{\sum (t^3 - t)}{12(N-1)} \right] \left[\frac{1}{n_i} + \frac{1}{n_0} \right]$$

where the sum is over all distinct responses and t is the number of observations tied at that response. The test statistic

$$Z = \frac{R_i - R_0}{\sqrt{V_{i0}}}$$

is compared to a standard normal distribution at $p=1-\alpha/k$, where k is the number of comparisons to control and α is 0.05 or 0.025, according as a 1- or 2-sided test is used.

This test is based on ranks, and thus is robust against a wide variety of distributions and heteroscedasticity. It is also flexible, so that arbitrary contrasts can be tested, not just comparisons of treatments to control. However, this latter is no advantage in a standard dose-response experiment. This test does not permit modeling multiple sources of variances (e.g., within- and between-subgroups). The Bonferroni-Holm adjustment for multiple comparisons is statistically conservative. There is no exact permutation counterpart to this test, so it is not useful for very small samples or experiments with massive ties among the response values.

Generally, Dunn's test is more powerful than multiple Mann-Whitney tests, but less powerful than Dunnett's test (where the requirements for that test are satisfied) and even less powerful than the Jonckheere-Terpstra test for data where the concentration-response is at least approximately monotone.

Dunnett (continuous response)– Definition and reference to the statistical test

The Dunnett test is based on T-tests and in fact the basic statistic is that for the Student T-test:

$$T = \frac{\bar{x}_i - \bar{x}_0}{s\sqrt{\frac{1}{n_i} + \frac{1}{n_0}}},$$

where s is the square-root of the usual pooled within-group sample variance. T is compared to the *1*- or *2*-sided upper α equicoordinate point of a k-variate t-distribution with correlations defined below, and $df=n_0+n_i+n_i \ldots +n_{k-1}-k$.

Unlike the T-test, Dunnett's test uses the correlations among the comparisons to adjust for both the number of comparisons of treatments to control and for the fact that each two such comparisons are correlated by virtue of having the control in common. That is, the differences M_i-M_0 and M_j-M_0 are correlated. No further adjustment to the p-values is needed. Dunnett's test is preferred to multiple T-tests with Bonferroni-Holm adjusted p-values because it is more powerful (i.e., sensitive) and preferred to multiple unadjusted T-tests because it controls the false positive rate at the nominal 0.05 level (or whatever significance level specified by the user). Dunnett's test assumes normally distributed data with homogeneous variances. This test ignores any trend in the concentration-response and instead compares each treatment group to the control without regard to how other treatment groups compare to the control or to each other.

Fisher exact test

Test for quantal data (i.e., count data, the number of affected organisms compared to the number of tested organisms). Fisher's Exact test is based on a 2x2 contingency table where control and a single treatment group are compared according to their prospective counts (Affected/Not affected). The diagram below illustrates this case.

	Control	Treatment	Total
Affected	n00	n01	n0.
Not Affected	n10	n11	n1.
Total	n.0	n.1	n..

Fisher's exact test is based on the probability of observing n_{01} affected subjects in the treatment group, if all marginal totals are considered fixed. This test considers only one treatment group and control. It makes no use of any dose-response characteristics observed. If the test is applied to all treatment groups, these tests are treated as though they are independent and some adjustment for the number of comparisons (i.e., treatments) should be considered, such as the Bonferroni-Holm adjustment to significance levels (p-values).

F-test (continuous response)- Definition and reference to the statistical test

The F-test is a general test for differences among group means. No consideration is given to whether a group is a control or treatment, so a significant F-test may mean there are two treatment group means that are different when neither is different from the control. Conversely, there can be a treatment group mean that is significantly different from the control mean and the F-test is not significant. This can happen because the F-test must "guard" against many comparisons of no interest in toxicology and so can miss an effect that is important in that field. Partly for that reason, a significant F-test should never be required in order to use one of mean comparison tests described below, which are attempts to focus attention on comparisons of interest to toxicology, namely comparisons of treatments to control. Requirements for the F-test are normally distributed data with homogeneous variances.

Jonckheere-Terpstra (continuous response)– Definition and reference to the statistical test

The Jonckheere-Terpstra test is a step-down or fixed-sequence test procedure that can be used in the same situations as Williams' test. The calculation of the Jonckheere-Terpstra test statistic is based on *Mann-Whitney counts*. These Mann-Whitney counts can be thought of as a numerical expression of the differences between observations in two groups in terms of ranks. The idea of the Jonckheere-Terpstra test is very simple. Order the observations from all groups combined, from smallest to largest. Decide, on biological or physical grounds, what direction (increasing or decreasing) the dose-response has.

For each two groups i and j, with $i < j$ and $d_i < d_j$, examine each pair (x_i, x_j) of observations that can be made, with x_i from group i and x_j from group j. Count the number, O_{ij}, of these pairs which follow the expected order, $x_i < x_j$ (for increasing trend; order reversed for decreasing trend). Add all of the O_{ij} and compare that sum to what would be expected if the dose-response were flat. A large positive difference is evidence of a significant increasing dose-response.

As a rank-based procedure, this test is robust against both mild and major violations of normality and homoscedasticity. There is an exact permutation version of the test available in commercial software (e.g., SAS and StatXact) to handle situations of small sample sizes or massive ties in the response values. It is based on a presumed monotone dose-response and is powerful against ordered

alternatives for a wide variety of dose-response patterns and distributions. There is no problem with unequal sample sizes if individual subjects are the experimental units for analysis.

While there is a common misconception that all non-parametric tests, such as the Jonckheere-Terpstra test, have inferior power to parametric tests, this is not universally true. In particular, for normally distributed, homogeneous data, the power properties of the Jonckheere-Terpstra and Williams' tests are very similar, with each being slightly more powerful than the other under some circumstances. Where the data are either not normal or not homogeneous, the power properties of the Jonckheere-Terpstra test are superior to those of Williams' test. These observations have been documented in extensive power simulation studies for many concentration-response shapes, number of treatments, and sample sizes.

However, there is no way to take into account multiple sources of variance, such as within- and between-subgroups. A consequence of this is that if the experimental unit for analysis is a subgroup mean or median and these subgroups are based on unequal numbers of subjects, then no adjustment can be made for this inequality. This is not an issue unless there is considerable variation in the number of organisms in different replicate vessels across concentrations.

As a step-down fixed sequence test, there is no need to adjust the p-values for the number of comparisons to control the false positive rate at 0.05 or other specified value.

Mann-Whitney (continuous response)– Definition and reference to the statistical test

The Mann-Whitney rank sum test compares the ranks of measurements in two independent random samples of n_1 and n_2 observations and aims to detect whether the distribution of values from one group is shifted with respect to the distribution of values from the other. It is equivalent to the Wilcoxon test.

To use the Mann-Whitney rank sum test, we first rank all $(n_1 + n_2)$ observations, assigning a rank of *1* to the smallest, *2* to the second smallest, and so on. Tied observations (if they occur) are assigned ranks equal to the average of the ranks of the tied observations. Then the ranks of the observations in each group are summed and designated as T_1 and T_2. If the distributions in the two groups are identical then T_1 and T_2 would be identical. If the two distributions differ, then the difference between T_1 and T_2 will be dissimilar, with the rank sums indicating the degree of overlap between the groups. There are one tailed and two tailed versions of the test, as well as small sample and large sample (asymptotic) versions.

The Mann-Whitney test is a non-parametric test that compares each treatment to the control without any reference to or use of data from other treatment groups. As a consequence, it has low power for small samples and is decidedly lower in power than a T-test. It can be used in the same circumstances, and with the same adjustments to p-values for the number of comparisons, as the T-test for data that do not satisfy the normality and variance homogeneity requirements of the T-test.

Rao-Scott Cochran -Armitage- Definition and reference to the statistical test

Test for quantal data (i.e., count data, the number of affected organisms compared to the number of tested organisms). In the chi-square, Fisher Exact, and Cochran-Armitage tests, there is no distinction made between organisms in the same replicate vessel or in different replicate vessels. Sometimes, when there are multiple organisms in each test vessel and multiple test vessels in each treatment group, there is more variability between replicates than simple binomial probability would indicate. This can happen when there is differential mortality among replicates and also when there is competition among subjects within a replicate, or for other reasons. In order to capture both within- and between-replicate variability for quantal responses, Rao and Scott modified the Cochran-Armitage test. Details are given in [Rao&Scott (1992), Rao&Scott (1999), Fung *et al*, 1994].

T-test (continuous response)

The T-test is a parametric test to compare a single group or treatment mean to another single group mean. In toxicology experiments, the only comparisons of interest are treatment to control. The error term used in a typical T-test is from an ANOVA involving all treatments and control. If a T-test is used to compare each treatment to control, then some adjustment for the number of comparisons (i.e., treatments) is advisable to control the false positive rate. The Bonferroni-Holm adjustment is often a good choice. A significant F-test is not required in order to test for treatment effects. Some of the reasons for this are discussed in the description of the F-test. Another reason is that the so-called protected F-test (i.e., comparing treatments to control only if the F-test is significant) changes the false positive and false negative rates of the T-test.

Tamhane-Dunnett (continuous response)

This test assumes normally distributed data but is optimized for variance heterogeneity. There is little power loss for homoscedastic data. The basic statistic is quite simple:

$$T = \frac{\bar{x}_i - \bar{x}_0}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_0^2}{n_0}}}$$

For a 2-sided test, T is compared to the maximum modulus distribution for k comparisons and $df=n_0+n_i-2$. For a 1-sided test, T is compared to the Studentized maximum distribution for same k and df.

If there are subgroups in the treatment groups, the sample variances in the above formula are replaced by Satterthwaite-type expressions. This test allows heterogeneous variances, but loses little power, compared to Dunnett's test, when the variances are homogeneous. It can be adapted to handle multiple sources of variance (e.g., within- and between-subgroups). This test ignores any trend in the concentration-response and instead compares each treatment group to the control without regard to how other treatment groups compare to the control or to each other.

Williams (continuous response)- Definition and reference to the statistical test

Williams' test is a step-down or fixed-sequence test procedure that can be used in the same situations as the Jonckheere-Terpstra test. Unlike the latter, Williams' is based on normally distributed, homogeneous responses and formally incorporates the presumed monotone dose-response in the estimated mean effects at each dose. These means are called isotonic estimates and are based on maximum likelihood theory, given the dose-response is monotone. Isotonic estimators were developed by Ayer *et al* (1955), who called their method Pool-the-Adjacent-Violators (PAVA) algorithm. Isotonic regression was introduced by Barlow *et al* (1972).

Where the concentration-response curve is at least approximately monotone, Williams' test is more powerful than pairwise methods such as Dunnett and Tamhane-Dunnett because it uses the additional information of a presumed monotone concentration-response. That is, it takes into account not just how a single treatment group compares to the control but how other treatment groups compare to the control and to each other. As a step-down fixed sequence test, there is no need to adjust the p-values for the number of comparisons to control the false positive rate at 0.05 or other specified value.

3.12 References

Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E (1955). "An Empirical Distribution Function for Sampling with Incomplete Information." The Annals of Mathematical Statistics, 26, 641–647.

Backhaus, T., Blanck, H., Faust, M. (2010) Hazard and risk assessment of chemical mixtures under REACH – State of the art, gaps and options for improvement, Swedish Chemicals Agency, Sundbyberg.

Barlow R.E., Bartholomew D.J., Bremmer J.M. and Brunk, H.D. (1972) - Statistical inference under order
restrictions, Wiley 1972.

Belden, J.B., Gilliom, R.J., Lydy, M.J. (2007) How well can we predict the toxicity of pesticide mixtures to aquatic life? Integr. Environ. Assess. Man. 3:364-372.

Chapman, P., Crane, M., Wiles, J., Noppert, F., McIndoe, E. (1996) Improving the quality of statistics in regulatory ecotoxicity tests. Ecotoxicology 5: 169-186.

Dhaliwal, B., Dolan, R., Batts, C., Kelly, J., Smith, R., Johnson, S. (1997) Warning: Replacing NOECs with point estimates may not solve regulatory contradictions. Environmental Toxicology and Chemistry 16:124-126.ECHA (2008), Guidance on information requirements and chemical safety assessment, Chapter R.10: Characterisation of dose [concentration]-response for environment, Available on http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r10_en.pdf?vers=20_08_08

Dunn O. J. (1964) - Multiple Comparisons Using Rank Sums, Technometrics 6, 241-252.

Dunnett C.W. (1955) - A multiple comparison procedure for comparing several treatments with a control, J. American Statistical Association 50, 1096-1121.

EU (1991) Council Directive 91/414/EEC of 15 July 1991 concerning the placing of plant-protection on the market.

EU(2009), Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC

Available on: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:309:0001:0050:EN:PDF

EU (2010), Guidance document on aquatic ecotoxicology, Sanco/11843/2010 (draft), rev. July 2010

Hutchinson, T.H., Ankley, G.T., Segner, H. And Tyler, C.R. (2006). Screening and testing for endocrine disruption in fish – Biomarkers as "signposts", not "traffic lights", in risk assessment. Environ. Health Perspect. 114: 106-114.

Kortenkamp, A., Faust, M., Scholze, M., Backhaus, T. (2007). Low-level exposure to multiple chemicals: reason for human health concerns? Environ. Health Perspect. 115 Suppl. 1:106-114.

Länge, R., Hutchinson, T.H., Croudance, C., Siegmund, F., Schweinfurth, H., Hampe, P., Panter, G.H., Sumpter, J.P. (2001) Effects of the synthetic estrogen 17α-ethinylestradiol on the life-cycle of the fathead minnow (*Pimephales promelas*). Environ. Toxicol. Chem. 20: 1216-1227.

OECD (2006) Current Approaches in the Statistical Analysis of Ecotoxicity Data: a guidance to application. OECD Series on Testing and Assessment. Guidance Document No. 54. Organisation for Economic Cooperation and Development, Paris, 146 pp.

OECD (2008) Detailed review paper on fish life-cycle tests, OECD Series on Testing and Assessment No. 95, ENV/JM/MONO(2008)22. Organisation for Economic Cooperation and Development, Paris.

OECD (2010) Short Guidance on the Threshold approach for Acute Fish Toxicity, Series on Testing and Assessment No 126, ENV/JM/TG(2010)/7, OECD, Paris .
http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=ENV/JM/MONO(2010)17&doclanguage=en

Oris, J. T., S. E. Belanger, and A. J. Bailer. (2012) Baseline characteristics and statistical implications for the OECD 210 Fish Early Life Stage Chronic Toxicity Test.Environ Toxciol Chem, 31(2), 370-6.

US EPA (1996) Federal Insecticide, Fungicide, and Rodenticide Act 7 U.S.C. §136 et seq., http://agriculture.senate.gov/Legislation/Compilations/Fifra/FIFRA.pdf.

US EPA (2004) Overview of the ecological risk sssessment process in the Office of Pesticide Programs, U.S. Environmental Protection Agency. Endangered and threatened species effects determinations. Office of Prevention, Pesticides and Toxic Substances, Office of Pesticide Programs, Washington, D.C., January 23, 2004.

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

Williams, T.D., Caunter, J.E., Lillicrap, A.D., Hutchinson, T.H., Gillings, E.G., Duffell, S. (2007) Evaluation of the reproductive effects of tamoxifen citrate in partial and full life-cycle studies using fathead minnows (*Pimephales Promelas*). Environ. Toxicol. Chem. 26: 695-707. 4. General test considerations.

4. GENERAL TEST CONSIDERATIONS

4.1 Concentration setting

99. Test concentration setting is an important part of the study design. Effective choice of concentrations ensures a successful test, minimizing the need for repeat runs, with consequent benefits in terms of animal welfare, time and cost. Key to the strategy employed is the purpose of the test in terms of quantitative endpoint determination (e.g. NOEC, LC_{50}) or qualitative hazard determination (e.g. endocrine screens to demonstrate or exclude *in vivo* endocrine activity). For quantitative tests, concentration selection is optimized to achieve the point estimate. However, for qualitative hazard identification, test concentration selection is driven by the need to test in the "concentration space" most likely to observe the hazard whilst not confounding the interpretation (e.g. inducing other unwanted effects that may result from systemic toxicity). This difference is acknowledged in the existing test guidelines by recommendations for test concentration spacing factors. For quantitative tests, the spacing factor typically should not exceed 2.2 to 3.2 (OECD TGs 203 and 210, respectively). The scientific rationale, justifying such an approach for test concentration spacing, is given in Doudoroff et al. (1951) and was further supported by the work of Sprague (1969). For the fish endocrine screens, a factor not exceeding 10 is recommended (OECD TGs 229 and 230), i.e. a larger concentration span.

100. Exceptions to the strategies employed above exist, where a concentration-response curve may not be pursued. An example is the recently published guidance on the use of the threshold approach for fish acute toxicity testing (OECD 2010). Essentially, the approach uses a limit test at a single threshold concentration determined by the results of *Daphnia* and algae tests. If no mortality is observed in the limit test, the fish acute value can be expressed as greater than the threshold value. However, if mortality is observed, a full concentration response test is triggered (i.e. the regular OECD TG 203). Another example of use of a single concentrations concerns chemicals of low toxicity, for which a limit test at the maximum concentration can be performed to demonstrate the absence of effects. Maximum test concentrations may be specified by the regulatory body. For example, for OECD tests it is 100 mg/L for acute tests (e.g. OECD TG 203) and 10 mg/L for chronic and endocrine tests (e.g. OECD TGs 210 and 229). However, other regulatory bodies may specify other maximum test concentrations, for example the maximum test concentration is 1000 mg/L for industrial chemicals (OPPTS 850.1075, U.S. EPA 1996).

101. Range finding tests, to inform definitive test concentration selection, are an important tool. However, the need for such confirmatory data should be weighed against the existing data on the test item or related substances. Existing data may negate the need for range-finding, since the effect range can be reasonably predicted. For example, effect levels can often be reliably predicted for formulated plant protection or biocidal products from existing data on the component active ingredient(s) that drive the overall toxicity. Mammalian and non-mammalian data for pharmaceutical studies may be useful, as could be read across from similar compounds (ECETOC 2007) for setting test concentrations. However, where such data are not available, it is often in the interests of animal reduction to run a well-designed range-finding test to avoid the need to repeat definitive tests that fail to capture the relevant endpoint(s).

102. Best practise for range-finding experiments is difficult to describe, as it is substance- and existing information-specific. Ideally, fish used in range-finding experiments should be as similar to the definitive test organisms as possible in terms of size and age (preferably from the same batch or source). For fish acute toxicity tests, typically three concentrations (no control) with a wide spacing factor of 10, when practical, and three fish per treatment offers sufficient information to set bounds around the approximate position of the LC_{50}. The rationale for this type of approach is given in Hutchinson et al. (2003). For chronic tests, range-finding is more difficult because of the longer duration, multiple endpoints and, due to reduced replication and differences in statistical power compared to definitive test. However, in general it is not necessary to range-find to the full duration of the definitive test, but only for sufficient time to assess the relevant parameters. For example, for the fish early life-stage test (OECD TG 210), duration of 14 days (ca. 9 days post-hatch in fathead minnow (*Pimephales promelas*) or 10 – 11 days post-hatch in zebrafish (*Danio rerio*)) usually allows for the fish to have grown sufficiently for estimation of growth effects. Again, a large spacing factor of 10 is useful. In general, range-finding should also be conducted in similar conditions to the definitive test in relation to exposure (static, semi-static, flow-through conditions). This is particularly important for unstable, volatile and high octanol-water partition coefficient substances. An appropriate internationally-accepted chemical analysis method should be available before the initiation of the range-finding study (see Chapter 4.5). Large differences in exposure between range-finding and definitive tests may occur due to differences in test systems and fish loading, resulting in range-finding that is not predictive.

103. For the endocrine screens, the purpose of range-finding is to ensure that the definitive test exposes fish to a suitable range of concentrations maximising the likelihood of observing the effect. The important distinction being that achieving a NOEC is not the purpose of the screening test, but rather to inform decision making for further testing investigations. Therefore, test concentration selection becomes a trade-off between testing sufficiently high concentrations to find the effect (if present) whilst not confounding the results by inducing systemic toxicity. To this end, the Maximum Tolerable Concentration (MTC), as defined by Hutchinson et al. (2009), is cited in the test guidelines OECD TGs 229 and 230. The MTC is the highest concentration that does not lead to a reduction in survival, feeding, normal behaviour and normal morphology and colour. Importantly, professional judgement is required to analyse all the available data to determine if the MTC can be estimated without range-finding. Here it is important to note that predicting the MTC from acute lethality data is problematic, since the screens are effectively long-term exposures (21 days for TG 229 and TG 230) assessing sublethal effects. Fish LC50 data from one timepoint alone are often insufficient for estimating the MTC. Therefore, chronic data are important in determining a suitable MTC. Furthermore, when range-finding is required, it is recommended that durations longer than acute tests be used. The test should assess indicators of systemic toxicity (mortality and symptoms of toxicity). These can then be used to estimate the MTC to be used as the highest definitive test concentration.

104. In summary, the purpose of the test needs to be considered in order to decide on an appropriate concentration range. In line with animal welfare considerations, where possible, existing information alone should be used to determine definitive test concentrations. However, it is often justifiable to conduct a suitable range-finding experiment to inform test concentration selection (leading to a higher test success rate and overall reduction in animal numbers).

4.2 Preparation of test solutions, including solvent-free methods

105. The preparation of test solutions is an important part of the experimental design. The physicochemical properties of the test item can make testing in aqueous media difficult. The Guidance document on aquatic toxicity testing of difficult substances and mixtures (OECD 2000)

provides useful guidance for such substances. This includes recommended organic solvents and solvent-free preparation methods. The preference is always to present the test item in the form it is most likely to occur in the environment, which for single chemicals is generally as the dissolved form, although preparations (pesticide formulations and other mixtures) may be tested as homogenous emulsions or suspensions. However, the physicochemical properties of the substance and necessary delivery options (e.g. flow-through systems) often mean it is not possible or practical to simply dissolve the substance in the test media. The objective is to achieve a biological (toxicological) response with testing up to the practical solubility limit in test media. It should be acknowledged that it is not always possible to achieve a biological response below the practical solubility limit and so a limit test at this concentration is often the best approach.

106. The most common practise to aid dissolution is the use of a solvent. However, this can be problematic in itself as it may potentially alter the bioavailability of the test substance and/or influence the test system (additional carbon source leading to microbial growth). As required by the individual test guidelines, a solvent control group should always be included with as many replicates as the water control group. Further, a recent review observed evidence that some low concentrations of solvents may affect the reproduction of certain fish species, and also impact biomarkers of endocrine disruption (Hutchinson et al. 2006). Therefore, where ever practically feasible, the use of solvents should be avoided.

107. Guidance for methods of solvent-free preparation has been described (Rufli et al. 1998, OECD 2000) and include generator columns, coating of stock solution vessels, sonication, and large volume (typically dilute) saturated aqueous stock solutions. Note, some of the non-solvent methods may result in the formation of micelles of the test substance, and this should be considered in test stock preparation and chemical analysis. However, it must be acknowledged that there are limitations to these methods, notwithstanding the considerable increase in time and cost associated with implementation in commercial and necessarily high throughput ecotoxicology laboratories. Difficulties also arise depending on the duration of the test and the supply of stock solutions, for example, to a flow-through delivery system. With these durations and at concentrations required, it may be difficult to ensure the stock concentration is maintained over the duration of the test. Ultimately, if not controlled, this may lead to unacceptable variability in test solution concentrations. Since maintaining acceptable variability in exposure solutions is a validity criterion (e.g. ± 20 % of the mean measured values; OECD TG 210), this has major implications for a laboratory's ability to conduct a valid study.

108. Therefore, there is a place for the use of solvents; however, it is sensible considering the issues described above to minimize the solvent concentration as is recommended in the current fish endocrine screening assays (OECD TGs 229 and 230). The maximum solvent concentration in chronic studies recommended by OECD is 100 µL/L (OECD, 2000), but Hutchinson et al. (2006) recommended, where solvent use is necessary in reproduction studies with aquatic organisms, the maximum solvent concentration should not exceed 20 µL/L of dilution water (0.002 %). This recommendation is a good target maximum concentration, although ultimately this may depend on the physico-chemical properties of the test substance, not least its solubility in solvents. However, where potential solvent effects are suspected, the potential influence of the solvent on the test results should be discussed (e.g. enhanced growth in the solvent controls). In summary, where ever possible, the use of a solvent should be avoided and alternative preparation techniques be employed. In cases where a solvent is required the concentration should be minimized, as far as practically possible.

4.4 Acclimation/culture maintenance/pre-treatment

109. The quality of test organisms is key to the successful conduct of fish tests. To ensure quality and confirm that the test organisms have adapted to laboratory conditions, the fish acute test (OECD TG 203) recommends a minimum acclimation period and batch selection criteria based on mortality and signs of disease during the acclimation period. In general, this is a practical method to ensure suitability for testing. For the longer-term studies, acclimation may not be possible (e.g. the fish early life-stage toxicity test starting with newly fertilised eggs). Here the preferred option is for embryos to be derived from in-house cultures of breeding fish where quality controls, such as not using a breeding group's first spawn (typically low viability) and disease status can be assured. However, it is acknowledged that this is not always possible. For example, it is not always practical to culture in-house all the required fish species (e.g. bluegill sunfish (*Lepomis macrochirus*)). Therefore, external suppliers such as commercial fish breeders are often necessary sources of test organisms. It is recommended that organisms supplied in this manner are accompanied by documentation from the supplier outlining basic information on source, occurrence of any treatments, time of collection (particularly for embryos) and any other pertinent information. This provides some safeguard against poor practise and establishes time lines for approximate time post-fertilisation, as required by the fish early life-stage toxicity test guideline (OECD TG 210).

110. For tests requiring reproductively active fish (OECD TGs 229 and 230), it is advisable to have a prolonged acclimation period. The time to a particular developmental stage of fish will differ, because of variability in certain parameters during their culture (e.g. feeding, temperature, density, etc.). Therefore, if it is not possible to culture the animals entirely in-house, a prolonged acclimation period will ensure they are fully adapted to laboratory conditions and are more likely to be at the appropriate developmental stage required in a test (e.g. actively spawning).

111. For some species, commercial sources may not be available, in which case the field collection of animals is required. In these cases, characterisation of the organisms and the site from which they are collected should be undertaken. Characterisation should include an assessment of the contamination history of the collection site, evidence that the animals are derived from a viable population (i.e. reproducing) and their parasite load. Once in the laboratory, acclimation of the population to laboratory conditions should include mortality, disease and stress assessment. Ideally, if the fish are to be used for endocrine screening (e.g. the androgenised female stickleback screen, AFSS), successful reproduction under culture/acclimation conditions is preferable before use in a test.

112. In summary, it is preferable for test organisms to be cultured in the testing laboratory. However, for certain test types and species this is not always practical. Information should be supplied with the batch of fish concerning their history. Strain should be included (e.g. OECD TG 210) where feasible; however, this may not always be possible for certain species. For the endocrine screening methods, longer acclimation periods may be necessary to ensure the fish are at the required developmental age/stage. Field collected species should undergo a full characterisation and acclimation assessment.

4.5 Species selection

113. Species selection considers a number of different factors including, size, ease of maintenance in the laboratory, convenience for testing, relevant economic, biological or ecological factors, known sensitivity, pre-existing data, animal welfare, availability of test methods for subsequent tests that may be triggered, as well as national or regional preferences, and especially for endocrine assays, validated endocrine biomarkers and endpoints of the species. There are also

practical considerations, such as the availability of cultured, as opposed, to field-collected organisms (see section 4.4 for considerations on acclimation/culture maintenance/pre-treatment). However, field-collected animals may be more appropriate for site-/situation-specific risk assessment questions. It is not always possible to meet all of these requirements within one test. However, species selection should always consider these factors, so further testing with additional species is less likely to be required.

114. In terms of acute toxicity, rainbow trout (*Oncorhynchus mykiss*) is considered to be amongst the most sensitive species. Dyer et al. (1997) reviewed sensitivity differences between tropical, temperate and coldwater species and found the latter consistently more sensitive for a diverse set of chemicals. This has also been established using acute toxicity Interspecies Correlation Estimation, a program developed by the USEPA, which indicate rainbow trout as more sensitive than fathead minnow, but often only at a factor of 2 to 3 (Dyer et al. 2006); see also chapter 5 on fish welfare in this document). Similarly, Lammer et al. (2009) summarized acute inter-species toxicity comparisons for zebrafish, Japanese medaka (*Oryzias latipes*), bluegill, fathead minnow and rainbow trout to 30 - 80 chemicals, depending on the species pair. Results were highly similar to those of Dyer et al. (1997, 2006). However, for fish early life-stage tests, smaller warm-water species (e.g. fathead minnow, Japanese medaka and zebrafish) are preferred, due to the shorter duration of the test, compared to the rainbow trout study (ca. 30 days *versus* ca. 90 days).

115. However, longer tests with rainbow trout may be sensible, when there are historical data on a class of compounds for which it advantageous to read across endpoints in the same species. For other tests species, preference may be driven by the endpoint of concern. For instance, rainbow trout is preferred in the fish juvenile growth test (OECD TG 215), since relative growth in the exponential phase is greater than in other species making differences easier to detect. For endocrine-specific testing, there are clear advantages of using the same species throughout general toxicity testing, endocrine screening and definitive endocrine testing. Such an approach could reduce the need for range-finding between test levels, i.e. general toxicity tests (e.g. fish early life-stage tests) could be used to set the MTC (see section 4.1 – concentration setting) for the fish endocrine screens (OECD TGs 229 and 230), all of which would inform concentration setting for definitive endocrine tests (tests currently available or under consideration are the fish sexual development test (FSDT – TG 234), the fish full life-cycle (FLCTT) and fish multi-generation tests).

116. Principles for selection of test species for chronic fish testing (e.g. OECD TG 210) would appear to be applicable to FLCTT testing (USEPA 1986, Benoit 1981), as well as to the FSDT. There should also be a consideration of whether the desired/necessary endpoints can be measured, or at least easily be measured in the species chosen. For example, if secondary sexual characters are a critical endpoint in either of the endocrine screens (OECD TGs 229 and 230), the species selected for testing should be Japanese medaka or fathead minnow rather than zebrafish. However, for other endpoints, fecundity measurements (fractional *versus* continuous spawners), egg or larvae success and body size, making blood sampling easier, may also be considerations. Similarly, animal reduction may be addressed by choosing species optimal for a particular endpoint. For example, fish tests that include sexual determination could be preferably performed with a species that can be genetically sexed (e.g. Japanese medaka), as opposed to one that cannot, would allow for the use of fewer animals, if other endpoints measured in the test are not driving the number of animals to be used.

117. In summary, there are a number of factors driving species selection, and it is not always possible to satisfy these within one single test species. There is also value in keeping some flexibility in test species choice, for example to allow for freshwater/estuarine testing or the use of a fish test species in which genetic sex markers may have been recently developed, etc. Coherent principles for

long-term studies, but flexibility in test species should be considered. However, there are clear advantages (where possible) in terms of animal reduction for using the same species at higher testing tiers.

4.6 Chemical analysis

118. An appropriate internationally accepted test method should be available for range-finding studies and before the initiation of the definitive test. The analytical method should cover the anticipated test concentration range. The validation should include an assessment of the recovery from test media (i.e. from spiked samples) and determination of the limit of quantification. The purpose of the analysis is to confirm exposure, and, as a minimum, analytical samples should be taken at the beginning and at the end of the exposure period from all treatments and control(s). Where appropriate, for longer-term studies, samples should be taken at weekly intervals. Additional samples should be taken from the test system or stock solutions, at the discretion of the study director, to investigate the impact of any failures to the test system (failure of dosing systems etc.). In general, it is recommended to take analytical samples from all replicates at the start and weekly intervals thereafter (OECD TG 229). However, for well-understood compounds, it may be scientifically justifiable to measure concentrations in fewer replicates at every sampling interval.

119. Replicates should be alternated, unless otherwise stated in the test guideline (e.g. OECD TG 229); so samples are not taken from only one replicate throughout the study. Samples should be taken at the approximate midpoint of the test vessel. At each sampling interval, it can be useful to take two samples; one for analysis and one for storage as a back-up. When poorly soluble materials are tested, the samples should be centrifuged or filtered prior to analysis and the supernatant analysed to determine the concentration of the test substance in solution, as this is presumed to be that which is biologically available. It should be noted that some of the non-solvent methods may result in the formation of micelles of the test substance and centrifugation is highly recommended prior to chemical analysis.

120. Ideally samples should be analysed immediately. However, often this is not practical and the samples must be stored (e.g. refrigerated or frozen depending on the test item) until they can be analysed. If stored, storage stability should be confirmed (i.e. prepared storage stability spikes in test media). Where solid phase or liquid extraction of the sample is required, this can be conducted before storage, as this will often enhance the stability of the sample. It is recommended to add an internal standard (preferably isotope marked) to the sample before the extraction procedure to be able to correct the chemical concentrations. The back-up sample can be analysed, if necessary, to confirm any results outside expectation. The backup sample can be especially useful for flow-through studies, where the dynamic nature of the test system means it would otherwise be difficult to investigate erroneous results. In conjunction with the chemical analysis, prior to the start of any flow-through study, the dosing system to be used should also be checked to confirm correct delivery of the test solutions. For flow-through studies, it is advisable to conduct pre-exposure analyses to ensure the test system is in equilibrium and test concentrations are approximately in the expected range, before adding test organisms.

121. In summary, suitable validated test methods should be available before initiation of the definitive test. Sampling should be at the beginning, end and at regular intervals during the exposure. However, the various regulatory authorities may have different analytical requirements.

4.7 Water and diet quality

122. Water and diet should be of sufficient quality to support normal test organism growth and development. This can be demonstrated by the culture of fish in the medium used for the test. The ability to culture test organisms through a life-cycle provides definitive evidence for appropriate water quality and culture conditions. However, for certain test types (e.g. OECD TG 203), this level of evidence is not required, as long as acclimation and test validity criteria pertaining to the biological quality of the organisms are met. Further, confirmation can be provided by the periodic chemical analysis of water (and sometimes food) for substances that may be present at levels considered to be toxic. The ASTM (2002) guidance can be consulted as a point of reference for this determination. There are also generally applicable criteria for water quality parameters in a recent OECD test guideline (OECD TG 229). Similarly, the nutritional value of the diet should be considered.

123. The clear advantage of commercial (formulated pellet or flake) foods is that nutritional quality is known. However, it can often be useful to provide additional sources such as frozen adult or brine shrimp (*Artemia* sp.) nauplii or other live food (*Daphnia*, *Chironomus* spp., etc.), particularly where fish reproduction (in culture or testing) is important. For fish endocrine tests (screening and definitive), the presence of potentially endocrine-active contaminants or food components (e.g. high phytoestrogen content) should be avoided.

124. In summary, there are also generally applicable criteria for water quality parameters in the ASTM guidance document (ASTM 2002) and recent OECD test guidelines (e.g. OECD TG 229, 234). It may also be important to consider the nutritional value and presence of contaminants in the diet.

4.8 Test acceptability criteria

125. The current test acceptability or validity criteria, specified in the fish test guidelines, include control mortality, dissolved oxygen concentrations and water temperature. Longer-term studies also include variability around analytical measurements and biological criteria (fertilisation success or spawning activity). These are important criteria and should be used in the assessment of data quality and the decision over whether to repeat a test. However, there should be some latitude and professional judgement to assess the likely impact of deviations from these requirements. Typically, up to 10 % mortality (or 1 out of 7 in the fish acute OECD TG 203) or species-dependent hatch and post-hatch mortality in early life-stage test (e.g. OECD TG 210) is allowed in the control group(s).

126. The level of control performance may affect the power of the test (see Chapter 3 on statistical considerations). In general, this should be considered an important criterion, since it pertains to the quality of the test organisms and factors of the test system that may cause significant stress sufficient to impact the reliability of the results determined.

127. The criterion related to dissolved oxygen ensures suitable conditions for the fish. However, for longer-term studies, judgement should be made as to the duration and magnitude of observations below the dissolved oxygen criterion of 60 % air saturation. For instance, the rapid instigation of aeration to bring the levels above guideline requirements may be acceptable, if the duration and magnitude were unlikely to have adversely impacted the results of the test.

128. Similarly, the requirements for water temperature between vessels at any one time (± 1.5 °C) can be challenging to meet, and again judgement can be applied to assess the potential impact.

129. Criteria based on analytical variability can be difficult to meet, particularly for flow-through studies which are more prone to occasional errors or drift of dosing systems and for "difficult" test substances. However, justifications for such exceptions should be fully described in the report.

130. Flexibility over biological criteria may be difficult, since they exist to ensure quality of the test organisms, but also to ensure there are sufficient individuals available at the end of a test for the determination of certain endpoints (e.g. growth). However, it should be acknowledged these criteria are often challenging to meet in certain species (e.g. fertilisation success in trout and larval survival to the free-feeding stage in zebrafish).

131. Duration of the study also enhances the challenge to meet the acceptance criteria. It should also be noted that when studies have numerous test acceptance criteria, even though test acceptability criteria are individually reasonable, the probability of random variation causing failure to meet at least one can be relatively high. Flexibility in interpretation of small deviations from acceptance criteria in these studies is recommended.

132. Minor statistical deviations from performance criteria should not be used to reject scientifically sound studies. For example, suppose the criterion for fertility of eggs is 95 %, but an individual study achieves 92 %. This is only a minor deviation, because over the course of time it is discovered that highly experienced laboratories can achieve fertility between 90 and 95 % and, in such a case, allowances should be made for studies that fall below the criterion to be acceptable.

133. Acceptance criteria should be taken as a holistic view and not just a tick-box exercise.

134. In summary, professional and scientific judgement should be applied to test acceptance criteria, as a consequence of deviation can be a need to repeat tests with consequent increases in animal use. However, all exceptions should be justified and the potential impact assessed and reported.

4.9 References

ASTM (2002) Conducting acute toxicity tests with fishes, macroinvertebrates and amphibians. Standard E729-96. Am. Soc. Test. Mat., 100 Barr Harbor Drive, West Conshohocken, Pennsylvania 19428.

Benoit, D.A. (1981) User's guide for conducting life cycle chronic toxicity tests with fathead minnows (*Pimephales promelas*). Environmental Research Laboratory – Duluth, Duluth, MN. EPA-600/8-81-001

Dyer, S. D., S. E. Belanger, and G. J. Carr. 1997. An initial evaluation of the use of Euro/North American fish species for tropical effects assessments. Chemosphere 35(11):2767-2781.

Doudoroff, P., Anderson, B.G., Burdick, G.E., Galtsoff, P.S., Hart, W.B., Patrick, R., Strong, E.R., Surber, E.W., van Horn, W.M. (1951) Bioassay methods for the evaluation of acute toxicity of industrial wastes to fish. Sewage Industr. Wastes 23: 1380-1397.

Dyer, S.D., Versteeg, D.J., Belanger, S.E., Chaney, J.G., Mayer F. L. (2006) Interspecies correlation estimates (ICE) predict protective environmental concentrations. Environ. Sci. Technol. 40: 3102-3111.

ECETOC (2007) TR 102 - Intelligent Testing Strategies in Ecotoxicology: Mode of Action Approach for Specifically Acting Chemicals. December 2007.

Fung,K.Y, D. Krewski, JNK Rao, AJ Scott (1994),Tests for trend in toxicity experiments with correlated binary data, Risk Analysis 14, 639-648.

Hutchinson, T.H., Shillabeer, N., Winter, M.J., Pickford, D.B. (2006) Acute and chronic effects of carrier solvents in aquatic organisms: A critical review. Aquat. Toxicol. 76: 69-92.

Hutchinson, T.H., Bögi, C., Winter, M.J., Owens, J.W. (2009) Benefits of the maximum tolerated dose (MTD) and maximum tolerated concentration (MTC) concept in aquatic toxicology. Aquat. Toxicol. 91: 197-202.

Hutchinson, T.H., Barrett, S., Buzby, M., Constable, D., Hartmann, A., Hayes, E., Huggett, D., Länge, R., Lillicrap, A.D., Straub, J.O., Thompson, R.S. (2003) A strategy to reduce the numbers of fish used in acute ecotoxicity testing of pharmaceuticals. Environ. Toxicol. Chem. 22: 3031-3036.

Lammer, E., Carr, G.J., Wendler, K., Rawlings, J.M., Belanger, S.E., Braunbeck, T. (2009) Is the fish embryo toxicity test (FET) with the zebrafish (*Danio rerio*) a potential alternative for the fish acute toxicity test. Comp. Biochem. Physiol. 149C: 196-209.

OECD (2000), Guidance document on aquatic toxicity testing of difficult substances and mixtures, Series on Testing and Assessment No. 54,t, OECD, Paris..

OECD (2010) Short guidance on the threshold approach acute fish toxicity, Series on Testing and Assessment No. 126, OECD, Paris.

Rao J.N.K. and Scott A.J. (1992) - A simple method for the analysis of clustered binary data, Biometrics 48, 577-585.

Rao J.N.K. and Scott A.J. (1999) - A simple method for analyzing overdispersion in clustered Poisson Data, Statistics in Medicine 18, 1373-1385.

Rufli, H., Fisk, P.R., Girling, A.R., King, J.M.H., Länge, R., Lejeune, X., Stelter, N.**, Stevens, C., Suteau, P., Tapp. J., Thus, J., Versteeg, D.J., Niessen, H.J.** (1998) Aquatic toxicity testing of sparingly soluble, volatile and unstable substances and interpretation of use of data. Ecotoxicology and Environmental Safety 39: 72-77.

Sprague, J.B. (1969) Measurement of pollutant toxicity to fish I. Bioassay methods for acute toxicity. Water Res. 3: 793-821.

US EPA (1986) Hazard evaluation division standard evaluation procedure – Fish life-cycle toxicity tests. EPA 540/9-86-137. United States Environmental Protection Agency, Washington DC.

US EPA (1996) OPPTS 850.1075 – Fish Acute Toxicity Test, Freshwater and Marine. EPA 712-C-96-118. United States Environmental Protection Agency, Washington DC.

5. ANIMAL WELFARE CONSIDERATIONS AND ALTERNATIVE APPROACHES

135. The purpose of this chapter is to provide background on current approaches to replacing, reducing and refining the use of animals in chemical testing. The first section describes social and legal impetus for reducing reliance on animal testing, while the rest of the chapter discusses general approaches to replace, refine or reduce animal testing, using examples in current use. Examples from ecological or fish testing are presented where available. Examples of the use of integrated strategies applied specifically to fish test guidelines will be presented in Chapter 7.

5.1 The "3Rs"

136. The concept of refining, reducing and replacing the use of animals (the "3Rs") in scientific endeavours was first articulated by Russell and Burch in 1959.[3] Although it is felt that animal testing is still necessary in toxicology, it is also a widely held view that animal suffering should be minimised and the use of animals be replaced or reduced where feasible. Especially in Europe, the "3Rs" are reflected in legislation on the protection of laboratory vertebrate animals in general and in legislation covering specific programs regulating safety assessment of pesticides, industrial chemicals and pharmaceuticals (see below).

5.1.1 OECD commitment to "3Rs"

137. The OECD has long stated its commitment to the "3Rs" and considers legally binding a Council Decision regarding the Mutual Acceptance of Data (MAD) as a main element of reducing the number of animals used in regulatory testing. MAD states that "safety data developed in one Member country will be accepted for use by the relevant registration authorities in assessing the chemical or product in another OECD country" and is intended to avoid duplicative testing. In addition, Test Guidelines (TGs) are periodically reviewed to update and harmonise animal welfare principles and/or reduce the number of vertebrate animals used.

138. Through the Task Force on Hazard Assessment, OECD member countries have also supported the development of the eChemPortal chemical information database (http://.webnet3.oecd.org/echemportal/) and the (Q)SAR Application Toolbox (http://www.oecd.org/document/54/0,3343,en_2649_34373_42923638_1_1_1_1,00.html), both of which can decrease vertebrate testing. Through the Joint Meeting, the OECD helps member countries pursue cooperation on chemicals testing workshops, initiatives, and programmes that support the development and advancement of the "3Rs."

5.1.2 European legislation

139. Principles of the "3Rs" are incorporated into several pieces of legislation in the European Union, such as the Council Directive 86/609/EEC (EU 1986) which specifically prohibits the use of animals where "another scientifically satisfactory method of obtaining the result sought, not entailing the use of an animal, is reasonably and practicably available" and also encourages EU Member States to develop alternative methods that adhere to the "3Rs". This Directive has been revised, and the new Directive 2010/63/EC eventually published on 20 October 2010 aims to

[3] Refinement: improvements that reduce pain and suffering considering the lifetime of the animal. Reduction: improvements in protocols that use fewer animals and or obtaining more information per animal. Replacement: methods that obtain biologically relevant information without the use of animals.

strengthen legislation and to improve the welfare of laboratory animals ((http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:276:0033:0079:EN:PDF; EU 2010).

140. A direct result of Council Directive 86/609/EEC was the creation of the European Centre for the Validation of Alternative Methods (ECVAM, http://ecvam.jrc.it/index.htm) hosted by the European Commission Joint Research Centre, Institute for Health and Consumer Protection, Ispra (Italy), which is responsible for the development and validation of non-animal testing methods. In addition, the European Commission is funding the development of non-animal alternatives for specific endpoints *via* its Framework Programmes (http://ec.europa.eu/research/fp7/)[4]. In addition to ECVAM, EU Member States have additional centres (such as ZEBET, NC3Rs, etc.) which are dedicated to the implementation of the "3Rs".

141. In the EU, the seventh amendment to the Cosmetics Directive represents unique provisions prohibiting the use of animals in testing of ingredients exclusively for use and marketed in cosmetics (EU 2003). This legislative initiative has translated into an economic incentive for industry to develop non-animal methods for common cosmetic tests. In 2009, a ban on animal testing within the territory of the EU went into effect for human health endpoint testing on chemicals that are exclusively used in cosmetics. In addition, a marketing ban went into effect on the sale of cosmetics that contain substances that have been tested on animals for all human health effects with the exception of repeated-dose toxicity, reproductive toxicity and toxicokinetics (the marketing ban applies to testing that has been carried out after the entry into force of the legislation: 2009 for short-term tests, 2013 for longer-term tests). Marketing of cosmetics containing substances that have been tested on animals for these endpoints is scheduled to be banned in 2013 (with the possibility that this decision may be changed e.g. the deadline may be postponed if non animal tests or non test approaches at that time are judged scientifically insufficiently developed and validated for fully replacing current Test Guidelines). Ecotoxicological testing of cosmetics is covered by Regulation (EC) No 1223/2009 on cosmetic products (EU 2009, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:342:0059:0209:EN:PDF) that comes into effect in 2013 and specifies that environmental concerns will be addressed under the REACH legislation. However, as regards chemicals only used in cosmetics and not in other chemical products, it is not yet clear how or whether REACH testing requirements on e.g. fish will be made consistent with the testing bans specified by the Cosmetics Directive, or vice versa.

142. The REACH legislation, adopted by the European Parliament in December 2006, contains both a high protection goal for man and the environment but also language and incentives to reduce the use of animals in testing (EU 2006). One of the primary aims of REACH is "the promotion of alternative methods for assessment of hazards of chemicals" (article 1.1), and the legislation stipulates that "information shall be generated whenever possible by means other than vertebrate animal tests" and animal testing shall be performed "only as a last resort" (article 25.1). Several

[4] For example, the Framework Program 6 (2002- 2006) was a collection of actions at the E.U. level to fund and promote research to replace animals for specific endpoints, including: PredictOmics to develop short-term *in vitro* assays for long-term toxicity, the ACuteTox project which has the overall objective of developing an *in vitro* test strategy that can replace *in vivo* testing for acute toxicity, ReProTect, focused on developing a testing strategy for reproductive/developmental toxicity, the OSIRIS (Optimized Strategies for Risk Assessment of Industrial Chemicals through Integration of chemicalsNon-Test and Test Information) project to develop integrated testing strategies (ITS) for REACH that will significantly increase the use of non-testing information for regulatory decision making and CAESAR (Computer-Assisted Evaluation of industrial chemical Substances According to Regulations) to create Quantitative Structure Activity Relationship models for the prediction of the toxicity of chemical substances [results from the final workshop can be found at http://www.caesar-project.eu/workshop/info.htm. (accessed 7 Sept 2010)].

other elements of REACH function to minimise redundant testing and, therefore, animal use, without compromising safety assessment. These elements include encouraging the formation of consortia of manufacturers or importers, grouping of chemicals (e.g. "read across" to fill data gaps), the use of (Q)SARs, the use of harmonized test guidelines including OECD test guidelines and integrated testing strategies. These various approaches are highlighted in a recent ECHA guidance document focused on "How to avoid unnecessary testing on animals" (ECHA 2010, http://echa.europa.eu/doc/publications/practical_guides/pg_10_avoid_animal_testing_en.pdf).

5.1.3 US legislation

143. Recent efforts to design more efficient regulatory testing schemes will result in the reduction of animal use, even if that is not a major objective. For example, the strategy outlined in the 2007 National Academy of Sciences report, Toxicity Testing in the Twenty-first Century: A Vision and a Strategy, calls for a strategy that progressively moves away from animal testing toward more mechanism-based molecular assessment (NAS 2007). This trend is also reflected in the US EPA's current Strategic Plan for Evaluating the Toxicity of Chemicals (US EPA 2009a, http://www.epa.gov/spc/toxicitytesting/).

144. This approach has been considered as part of proposed legislation to update the Toxics Substances Control Act 1976. Bills in both Houses of US Congress included a section addressing the minimisation of animal use and language encouraging the use and further development of approaches to minimise animal use (US Senate 2010). It is clear that the "3Rs" will be increasingly important in future regulatory considerations, on economic and practical as well as humane grounds.

145. Intended to facilitate the coordination of efforts to implement the "3Rs" in chemicals testing programs in the US, the Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM) was created within the National Institute of Environmental Health Science as a standing committee in 1999 (http://iccvam.niehs.nih.gov/). ICCVAM participates in international validation efforts of alternative methods and makes recommendations for use of validated methods to federal agencies for acceptance.

5.1.4 Initiatives to implement the "3Rs" in other countries

146. Following the example of the EC, Japan and South Korea have also created Centres for the Validation of Alternative Methods (JaCVAM, and KoCVAM, respectively) to facilitate incorporation of the "3Rs" into national testing programs and to participate in international efforts to validate alternative methods. JaCVAM was created in 2007 as part of the Japanese National Institute of Health Sciences; KoCVAM was created in 2010 as part of the National Institute of Food and Drug Safety (NIFDS) in the Korean Food and Drug Administration.

147. A Memorandum of Understanding has recently been signed creating the International Cooperation on Alternative Test Methods (ICATM) between ECVAM, ICCVAM, JaCVAM and the Environmental Health Science and Research Bureau within Health Canada. ICATM is designed to facilitate and harmonize the application of the "3Rs" in testing programs internationally.

5.2 Current approaches to testing frameworks

148. While different regulatory programs, countries and regions may employ different strategies to assess the risk of chemical exposure, basic elements of risk analysis are common to all, including specific testing protocols for hazard assessment. Advantages of applying generic frameworks in both hazard and risk assessments include the following:

- defined strengths and weaknesses of each test method;
- more efficient information evaluation;
- consistent evaluations and conclusions from information derived from different sources;
- and facilitated harmonization of regulatory acceptance of evaluations.

5.2.1 Tiered Testing Frameworks

149. One such generic framework is a tiered system of testing, consisting of sequential tiers (or batteries) of tests. Initial tiers generally consist of physicochemical data, read-across, chemicals categorisation, quantitative structure-activity relationship (QSAR) modelling, and *in vitro* characterization, followed by tiers of increasingly animal- and resource-intensive tests. Tiered testing has been applied to mammalian risk assessment, for example in the US EPA's Voluntary Children's Chemical Evaluation Program (US EPA 2000) and is for some endpoints included or encouraged under REACH (ECHA 2008a). The advantage of this type of tiering system is that a chemical assessment may be terminated following testing of each tier, depending on results.

150. REACH provides an example of another type of tiering system. In principle, REACH testing requirements are connected to production or import volume of chemical substances. However, there are several options for waiving or adaptation of the required testing on fish under REACH relating to annual tonnage marketed per manufacturer or importer, release potential, fate related properties of the substance and outcome of the chemical safety assessment. For example, the short-term fish test (OECD TG 203) for chemicals manufactured or imported into the EU at greater than 10 tonnes per year (Annex VIII) is not required, if a long-term aquatic toxicity study on fish is available or if there are mitigating factors indicating that aquatic toxicity is unlikely to occur (e.g. the substance is highly insoluble in water or is unlikely to cross biological membranes[5]). Other additional fish tests (long-term fish testing, either the OECD fish early life-stage OECD TG 210, the fish, the OECD short-term toxicity test on embryo and sac-fry stages OECD TG 212, or the OECD fish juvenile growth OECD TG 215) may also be required for substances at greater than 100 tonnes per year (Annex IX), and in some cases for substances below 100 tonnes/year, if the chemical safety assessment indicates the need to investigate further effects on aquatic organisms. On the other hand, registrants should consider long-term testing on fish for substances manufactured or imported in quantities of 10 tonnes or more, if their substances are poorly water soluble (Column 2 of Annex VIII 9.1.3.). In any case, the application of Tiered Testing Frameworks reduces the number of test animals used under REACH.

5.2.2 Integrated Testing Strategies (ITS)

151. Integrated Testing Strategies provide a more efficient framework that facilitates risk assessment of a large number of chemicals and relies on hypothesis-driven approaches. Integrated Testing Strategies employ all existing information, such as exposure information,

[5] In the "REACH Endpoint Specific Technical Guidance Document to Industry on the Information Requirements for Endpoint Specific Guidance on Aquatic Toxicity" (ECHA 2008, (p. 42/348)) it is stated regarding these adaptation rules that: "There is no scientific basis to define a cut off limit value for solubility below which no toxicity could occur"… and "Equally, there is no scientific basis to define molecular characteristics that would render a substance unlikely to cross biological membranes". For poorly soluble substances adaptation of the standard testing requirement should "be carefully justified and instead of an acute test it should be considered to perform a long-term test"

QSAR model predictions, chemical categorisation, threshold approaches, read-across, data from non-vertebrate test species, and *in vivo* and *in vitro* methods, and can be applied to any testing approach including tiered systems. Integrated Testing Strategies employ a decision tree, whereby at several stages in each scheme a decision is made, *via* a weight-of-evidence analysis, whether the information is sufficient to make a decision, or more testing is needed (Grindon et al. 2008; Fig. 5.1) taking into account the particular case in question and that different kind of decisions may not require the same type of information or the same level of uncertainty. The scheme also indicates what the next test could be. Ultimately, the goal is to utilize predictive approaches to identify what *in vivo* information is most relevant and necessary (Bradbury et al. 2004). There are numerous methodologies currently in use (such as those mentioned above) in hazard and risk assessment; however, there is a need to better integrate these approaches into true Integrated Testing Strategies frameworks (van Leeuwen et al. 2007).

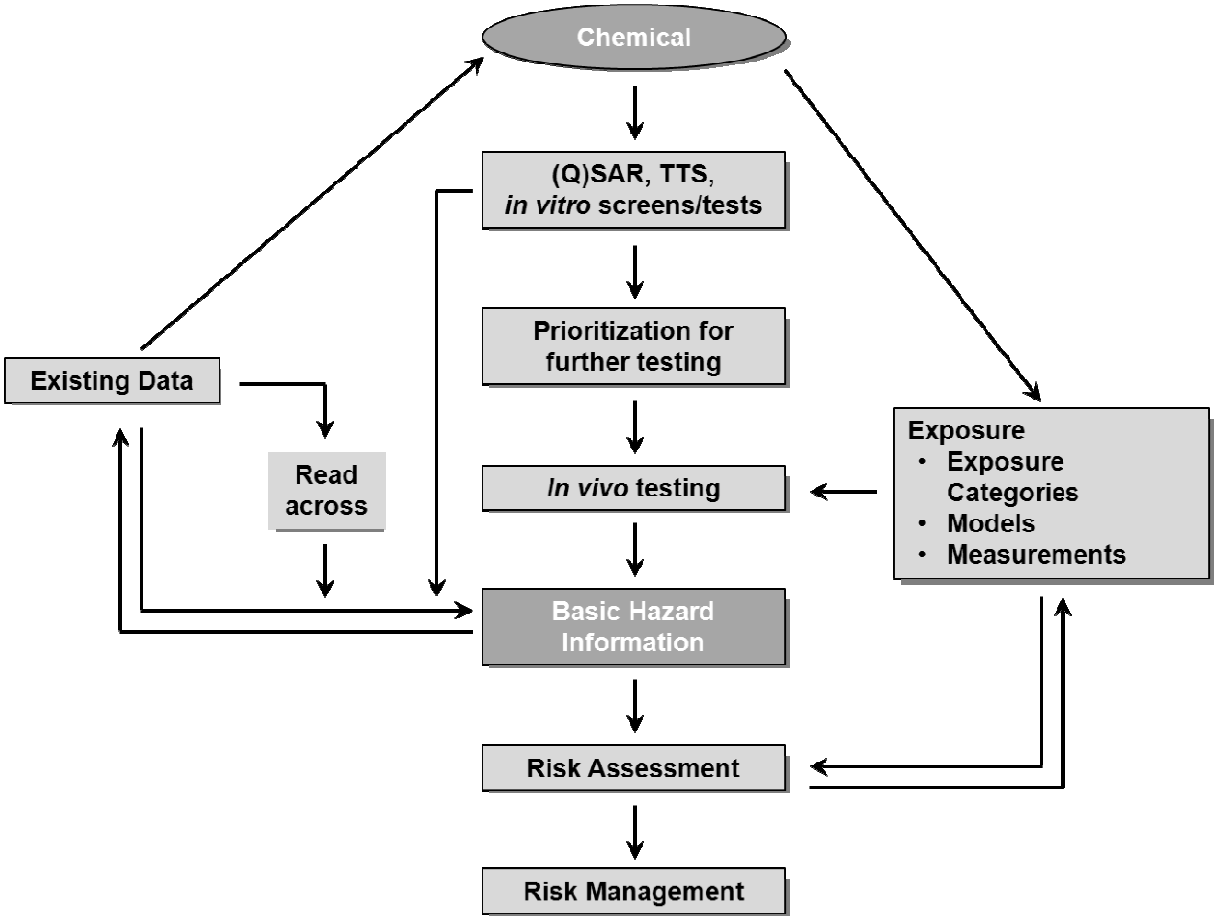


Fig. 5.1: Integrated Testing Strategy generic structure (from van Leeuwen et al. 2007)

Note: TTS = Total Toxicity Score

152. The European Chemicals Agency (ECHA) has provided extensive guidance on information requirements and chemical safety assessment that includes Integrated Testing Strategies for specific endpoint assessment, including ecotoxicological endpoints. For example, Chapter R.7b of ECHA (2008a,b) includes an Integrated Testing Strategy for assessing acute aquatic toxicity, and

Chapter R.7c of ECHA (2008a,b) includes Integrated Testing Strategies for aquatic bioaccumulation, terrestrial bioaccumulation, and avian toxicity (ECHA 2008a, b)

153. Several other groups have proposed Integrated Testing Strategy frameworks for ecotoxicity testing. A review of proposed Integrated Testing Strategy approaches for aquatic toxicity is provided in Netzeva et al. (2007). It should be noted that discussion of particular methodologies that can serve as the components of an Integrated Testing Strategy are highlighted in the sections below.

154. An example of an Integrated Testing Strategy applied to bioaccumulation has been developed *via* an European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) workshop and described by de Wolf et al. (2007; Fig. 5.2).

155. The Integrated Testing Strategy described by de Wolf et al. (2007) involves a tiered process; tier one involves estimation models and tier two includes non-animal experimental systems. Depending on the properties of the chemical being evaluated and the resulting quality of the prediction, these tiers can lead to a complete replacement of animals used for assessing bioconcentration. Tier three includes animal testing, but with a reduced number of animals relative to the full OECD TG 305, which is reserved for tier four (the final step in the strategy). A weight-of-evidence (WOE) approach similar to this Integrated Testing Strategy is included in the ECHA guidance for REACH (ECHA 2008b).

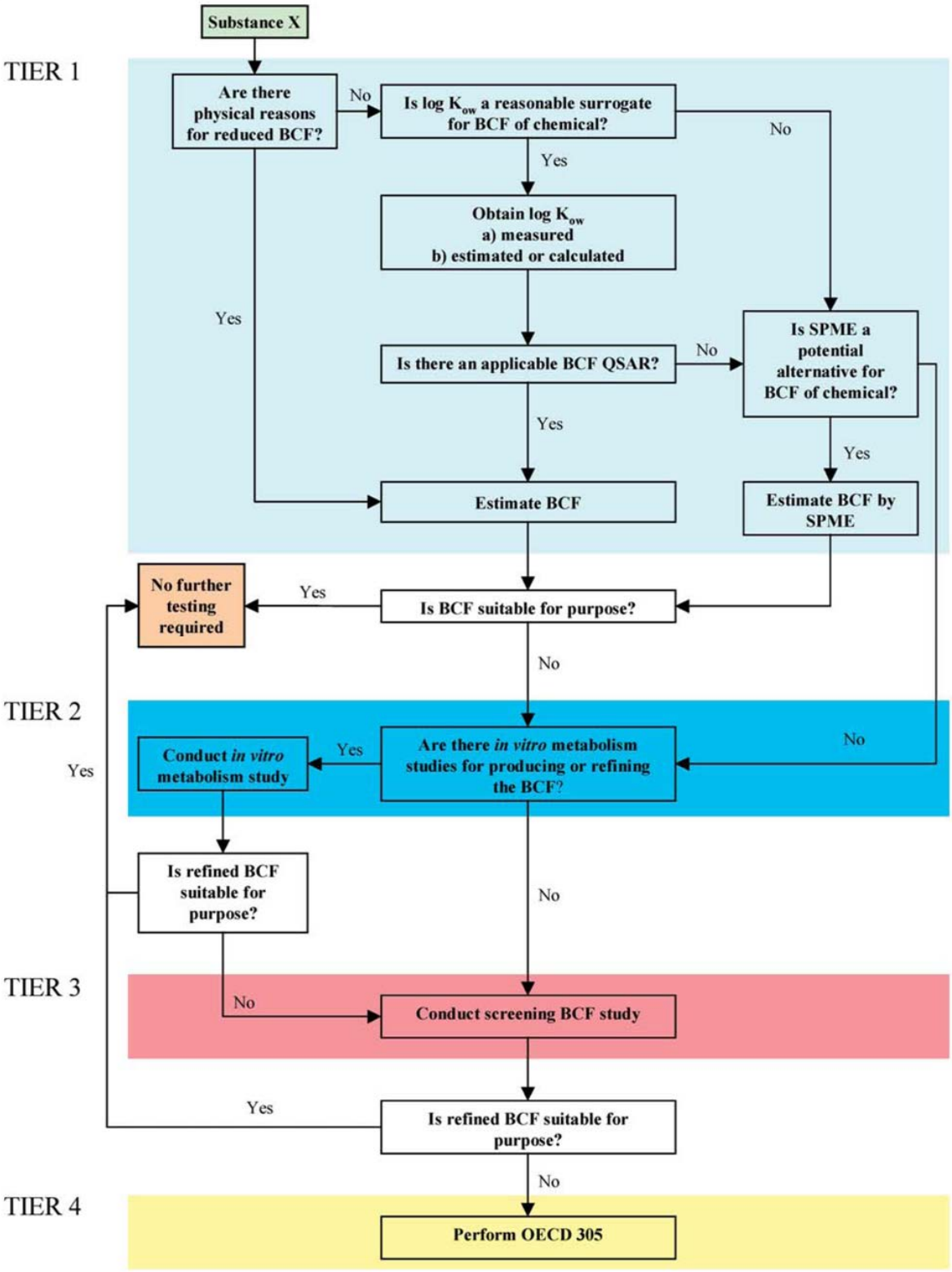


Fig. 5.2: Integrated Testing Strategy to assess fish bioconcentration (from de Wolf et al. 2007)

Note: SPME = Solid-phase microextraction

156. There is on-going work in various arenas to enhance models (Arnot et al. 2009), develop novel *in vitro* approaches (for review, see Weisbrod et al. 2009), and to revise the existing OECD TG on bioaccumulation in fish. This revision includes measuring BCF via uptake from water according to the current TG; also an option to reduce the number of fish used in the existing *in vivo* BCF test (OECD TG 305) (Springer et al. 2008); and another option suitable for very hydrophobic substances to measure BAF after dietary uptake. It should be noted that the Integrated Testing Strategy presented in Figure 5.2 does not take into account potential problems relating to testing and assessment of very hydrophobic substances for which the current OECD TG 305 is not suitable (due to problems by maintaining and measuring very low concentrations in the water etc.) The current revision of TG 305 does address this problem.

157. An overall Integrated Testing Strategy for ecotoxicology focused on reducing and replacing fish and amphibians in toxicity testing, was proposed by the European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) (ECETOC 2007, Hutchinson 2008). ECETOC reviewed several case studies assessing aquatic toxicity and bioaccumulation of several chemicals considering "mode of action" (MOA)[6]; in this case, MOA was defined as "a common set of physiological and behavioural signs … that characterize a type of adverse biological response." This study focused on the most highly active MOA category, that of specifically-acting chemicals, and divided this group into four subcategories based on type of protein interaction: receptor, ion channel, enzyme or transporter. The study concluded the need to utilize all available information on a chemical, including mammalian toxicity data, to provide insight into a chemical MOA which will help target testing to the most appropriate and sensitive species and identify specific tests for further characterization. This approach points out the potential future usefulness of genomics, proteomics and biomarker assessment for (i) setting doses for testing, (ii) determining the appropriateness of read-across or (iii) determining NOAELs. In addition, the "omics" and biomarker information could be used to define mechanism of action (specific biological activity), to identify critical toxicity pathways, and ultimately, to the extent possible, to design appropriate *in vitro* assays.

158. Critical improvements in development of Integrated Testing Strategies for human health and ecotoxicological endpoints is the focus of the Optimized Strategies for Risk Assessment of Industrial Chemicals through Integration of Non-Test and Test Information (OSIRIS), which is included in the Sixth Framework Programme Project funded by the European Commission. A detailed Integrated Testing Strategy for an overall assessment of fish toxicity under REACH has been developed by Roncaglioni et al. (2009) as part of the OSIRIS project. This scheme optimizes decisions for Classification and Labelling, Persistence, Bioaccumulation and Toxicity (PBT) Chemical Safety Assessment according to REACH requirements. In addition, an important objective of the OSIRIS project is to develop a generic strategy for Integrated Testing Strategies that includes quantitative estimates of certainty; an example of this approach using a simple three test system has been developed by Jaworska et al. (2010).

[6] Instead of referring to MOA, it could be considered to refer to AOP (Adverse Outcome Pathway), which is referring to MOA and also ADME

5.2.3 Additional strategies which, if incorporated into testing programs, result in reduced animal use

- Weight of evidence: All existing data is evaluated before further testing is proposed to assess (1) whether additional testing is necessary for the purpose at hand and, if so, (2) how to target that testing to provide the information needed. This is an element of Integrated Testing Strategies, but can be used on its own in certain circumstances (e.g. classification and labelling in some regulatory contexts). The ECHA technical guidance contains this description of weight of evidence: "It points to the likely properties of a substance. This approach may be applied, if there is sufficient information from several independent sources leading to the conclusion that a substance has (or has not) a particular dangerous property, while the information from each single source alone is regarded insufficient to support this assertion (see Annex XI, 1.2 of the REACH regulation for more detail)" (ECHA 2010). It is noted that use of all existing data including non-test information may in most cases but not always reduce animal testing. In some cases such available data may not be conclusive but nevertheless raise enough concern to trigger testing.

- Grouping chemicals based on structural similarities relative to intrinsic properties of chemicals (formation of chemical categories)

- Formation of consortia of companies facilitates the exchange of toxicological information to reduce testing needs.

- Public consultation of testing proposals, ensuring that the best possible use is made of existing information, and that animal testing is performed only when necessary.

- Development of validated *in vitro* methods, QSAR models and creation of chemicals categorisation tools which allow generation of transparent documentation of predictions or estimations including information about applicability domain (such as the OECD QSAR Application Toolbox)

159. Specific strategies for assessing aquatic toxicity within the OECD test guidelines programme is the subject of Chapter 7 of this document.

5.3 Optimisation of *in vivo* data

160. An approach to minimising animal use is the concept of maximizing the amount of information obtained from each animal used. An example of how to better use ecotoxicity data would be to calculate incipient LC_{50} for bioaccumulative substances where LC_{50} decreases versus time but does not reach the incipient asymptotic LC_{50} value for the substance in question.

5.4 Approaches for minimising fish use in acute toxicity testing

5.4.1 Range-finding

161. In situations where range-finding in concentration setting is needed, a significant reduction of animals in the definitive test may be obtained through the estimation of testing concentrations by use of strategies described in the following sections. The quality of range-finding may have a considerable influence on the number of animals needed in the definitive test and on the abundance of test repetitions. All options for range-finding without using animals should be considered:

- QSAR predictions, estimation from chemical categorisation, and read-across should be used, if reliable results are expected.

- For acute fish tests, the Fish Embryo Test (FET) gives an estimation of the range of the definitive LC_{50} (Lammer et al. 2009). It can be used in the approaches discussed below by requiring only a confirmatory reduced test design for the definitive test with fish.

5.4.2 Limit test

162. The concept of a limit test has been incorporated into most acute toxicity protocols, involving the use of a single group of fish to a pre-determined dose or concentration above which the chemical would be considered non-toxic (e.g., 100 mg/L in OECD TG 203). If no death occurs, no further testing is required. This approach is described in detail in the OECD TG 203 and is well-established in acute toxicity testing.

5.4.3 Threshold approach

163. The threshold approach describes a testing strategy which has the potential to significantly reduce the number of fish to be used for acute fish toxicity testing. It is based on the observation that fish is not always the most sensitive of the three testing species (fish, algae and invertebrates) generally used for short-term aquatic toxicity testing (Weyers et al. 2000, Hutchinson et al. 2003, Jeram et al. 2005).

164. As an initial step, an acute fish test is performed at a single concentration following the limit test method described in OECD TG 203. The single concentration (threshold concentration) corresponds to the lowest EC_{50} value from reliable algae (use the lower of the ErC50- and EyC50-values) and acute invertebrate (e.g. *Daphnia*) toxicity data as described by Hutchinson et al. (2003). If no mortality occurs in the limit test using the threshold concentration, it demonstrates that fish is not the most sensitive species at short-term exposure, and it can be concluded with 99 % confidence that the L_{C50} in fish is greater than the threshold concentration. If mortality is observed, a full OECD TG 203 stu*dy shou*ld be conducted.

165. The concept of the threshold approach is integrated into the REACH testing strategy for acute fish toxicity testing (ECHA 2008a) and has recently been approved as OECD Guidance Document no. 126 Short guidance on the threshold approach for acute fish toxicity (http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono%282010%2917&doclanguage=en).

5.4.4 Other approaches

Rufli and Springer approach

166. A new approach to reduce the numbers of fish required for the acute toxicity test (OECD TG 203) has been proposed by Rufli and Springer (2011). The approach is based on an extensive analysis of historical data from two databases (industry and the US EPA Office of Pesticide Programs) and simulation models. The models compared the current OECD TG 203 design (concentration range-finding: 4 fish, spacing factor 10; plus definitive test: 5 concentrations with 7 fish each, spacing factor 1.6) to study designs containing 4 or 5 test concentrations, 5, 6, or 7 fish in the definitive test, and 2, 3, or 4 fish in the range-finding study. The use of only four test concentrations results in a lower quality of the LC_{50} estimate for low-slope scenarios. Using six fish per concentration should yield LC_{50} estimates that are of a quality similar to those obtained using the seven fish presently required by OECD TG 203.

Sequential and step-down approach

167. The concept of the Up/Down approach is to sequentially expose a single or a few individuals at one concentration at a time, adjusting subsequent exposures higher or lower depending on the results. In this way, an LC_{50} can be estimated (*via* application of an algorithm) using fewer animals than by performing a complete dose-response at one time; a disadvantage of this approach is that the slope of the curve surrounding the LC_{50} is not established. It has also been used successfully in fish acute toxicity testing, but only to a limited extent because of the lack of regulatory acceptance.

168. Although not supported by OECD member countries another variation of the threshold approach described above was initially proposed. As described by Hutchinson *et al*. (2003) and also Jeram *et al* (2005), a fish test is performed at the threshold concentration using 5 test and 5 control fish. When no death occurs, further testing is not needed and the LC_{50} for fish is greater than the threshold concentration, as described above. When mortality is observed, further testing takes place at progressively lower concentrations until no mortality is observed, rather than a full LC_{50} test according to OECD TG 203. The LC_{50} would be calculated using the data from the tested concentrations.

5.5 Other considerations for *in vivo* testing

5.5.1 Animal welfare considerations for current test guidelines

169. Review and evaluation of existing *in vivo* test guidelines in light of the "3Rs" should be undertaken to ensure that existing test methods account for animal welfare considerations. One such test guideline that is often cited is the short-term toxicity test on embryo and sac-fry stages (OECD TG 212). This test has been termed by some as the "fish starvation test" due to the duration of the test post-hatch and the lack of external food supply provided to the test animals throughout the test. This test states that it is terminated, "…just before the yolk sac of any larvae in any of the test chambers has been completely absorbed or before mortalities by starvation start in controls" (OECD 1998).

5.5.2 Non-lethal endpoints

170. Available approaches that can be used to reduce animal suffering, such as reliance on more humane endpoints, are a key consideration in any refinement approach. Using "moribund" as the endpoint rather than "death" is seen as a refinement to reduce animal suffering, such that the test animals are humanely killed upon exhibiting toxic symptoms and considered unlikely to survive (ECETOC 2005) (OECD 2000). However, the definition of death has not yet been changed in the acute fish test (OECD TG 203), and the impact of using moribund as the endpoint instead of death might affect the magnitude of the LC_{50} values (e.g., make them lower). Based on a retrospective analysis of 328 fish acute toxicity tests of one laboratory and 101 tests of ten other laboratories from Europe and the United States (Rufli, 2012), the LC_{50} was lowered (more toxic) in up to 52% of the studies when moribund instead of death was used as the endpoint. The toxicity increase generally was by a factor of about 2, the maximum increase by a factor of 16. The period of suffering of the fish was reduced by 24 to 72h. To produce comparable results between laboratories when moribund is used as a sub-lethal endpoint requires the following specifications in a guidance document: 1) A unique definition of the moribund state in fish, 2) types of abnormalities and degree of effects to be reported.

5.5.3 General principles for minimisation of animal use within existing Test Guidelines

171. In an *in vivo* test, all options to reduce the number of fish should be evaluated, provided that they do not compromise the statistical power (and therefore usefulness) of the test. The starting point for this evaluation is the biological relevance of changes observed for a particular test endpoint, and this biological relevance should define the necessary precision of the measurement which then will define the necessary statistical power. Precision should be defined on the basis of biological considerations. Only then, the necessary group size and number of concentrations can be calculated by the statistical methods (see chapter 3 for more detail). Options for a reduced test design in a guideline should be very clearly defined (under which circumstances are they applicable?) to enhance the chance for general acceptance and thus for regular use. Some examples and considerations related to the minimisation of animal use within existing *in vivo* test guidelines are provided below:

- Test conditions: In fish tests, the variability of measured endpoints is often driven by environmental conditions, e.g. fish growth is highly dependent on temperature (even within the range specified in the test guidelines), food quality, and the available volume of water per fish. A closer definition of relevant environmental test conditions may therefore reduce the variability in the test and the necessary number of fish to reach a defined statistical power.

- Number of control fish: There is the need to consider whether both water and solvent controls are needed for various tests and whether one can rely on historical control values to reduce animal numbers.

- Evaluation of existing information and reduced design: Utilization of existing information on a chemical might provide insight into whether the results of a conducted *in vivo* test will prove relevant to the hazard or risk assessment context at hand and could lead to reduction in the numbers of animals utilized. For example, estimation of a bioconcentration factor (BCF) for organic substances can often be performed using advanced QSAR models or even by simple QSAR models based largely on log K_{OW} values. If this estimated BCF falls well below the trigger values used in hazard and/or risk assessment but testing is generally required, it should be considered depending on the predicted BCF value compared with regulatory relevant BCF cut off values whether a minimised test design with a markedly reduced number of fish with high probability would be sufficient to provide enough statistical power to confirm the expected value.

5.6 Species extrapolations (SSD, ICE)

172. Methodology to extrapolate toxicity data from test organisms to the organisms that must be protected is a constant source of debate, even in the human health arena where there is a single species to protect (human) and the vast majority of tests are performed in mammals (usually rodents). Ecotoxicity testing represents a much more complex scenario, where protection goals include all animal and plant species and tests are conducted on a small number of species.

173. The US EPA has developed a tool (ICE – Interspecies Correlation Estimation) to estimate a chemical's acute toxicity (LC_{50} or EC_{50}) to a species, genus, or family from the known toxicity of the chemical (provided by a database of acute toxicity values) or from an available estimate of toxicity (such as QSAR) to a surrogate species (Asfaw et al. 2003, Raimondo et al. 2010). The tool was subsequently internet-enabled (Web-ICE, Raimondo et al. 2010; see http://www.epa.gov/ceampubl/fchain/webice/index.html) with modules to predict acute toxicity to aquatic organisms (fish and invertebrates) as well as wildlife (birds and mammals). ICE models currently available on the web have been validated using leave-one-out simulations, sensitivity

analysis and detailed hand-checks of raw data, and thus are different than and improved from the initial models published by Asfaw et al. (2003).

174. Some generalities have been already discussed in the literature regarding how to consider use of ICE models. Models are sensitive to taxonomic distance, thus fish predict fish better than fish predict invertebrates, for example. ICE models appear robust even though they contain chemicals with diverse modes of action (MOAs). It is not known why this is so at this time, but the US EPA has active research on-going to develop ICE models for some MOAs having sufficient data, which could improve some models further (M.G. Barron, US EPA, pers. comm.). Plant (algae) data are largely missing, which is also being addressed (S. Belanger, P&G, pers. comm.). Marine toxicity data is less plentiful and often results in models that perform less well, but this could be addressed through additional input values. Lastly, the database is dominated by North American taxa; however, it is unlikely that the principles are any different when incorporating non-North American species. Lammer et al. (2009) provided ICE models for Japanese medaka (*Oryzias latipes*) and zebrafish (*Danio rerio*) to fathead minnow (*Pimephales promelas*), bluegill sunfish (*Lepomis macrochirus*), and rainbow trout (*Oncorhynchus mykiss*) and found good correlations.

175. A hazard and risk assessment option that is also web-enabled in Web-ICE is a module to generate Species Sensitivity Distributions (SSDs) from Web-ICE (Dyer et al. 2006, 2008). An SSD provides an estimated concentration predicted to be protective of some *a priori* level of species, most often 95 % (the so-called HC_5, EU TGD 2008; Stephan et al. 2002). The tool uses limited input data (say a single fish acute toxicity value) to predict other fish and invertebrates, whose collective toxicity output is subjected to SSD analysis to generate an HC_5. Dyer et al. (2006, 2008) used this concept to estimate hazardous concentrations of equal quality to US EPA ambient water quality criteria. In theory, high quality QSARs could be used to predict fish toxicity to an unknown chemical (as well as invertebrate and algal toxicity) and provide hazard estimates that would not require any test animals and still be predictive enough to not apply very large uncertainty factors (Belanger et al. 2009).

176. The above described cross species estimations use acute data alone but may be useful in relation to priority setting of testing needs when no other measured or predicted information can be obtained.

5.7 QSAR methods

177. Quantitative Structure Activity Relationship (QSAR) models are increasingly viewed as one of the most cost-effective ways to estimate ecological and health effects of chemicals, though most of the current use by regulatory agencies is on priority setting for existing chemicals and classification for chemicals, and to identify needs for further testing and/or testing strategies. However, it should be noted that QSAR predictions for acute toxicity and bioconcentration in fish are among the most commonly used and well accepted for regulatory purpose. Several models exist for estimating bioaccumulation, and especially bioconcentration in fish. QSAR models have also been developed that are useful for predicting acute aquatic toxicity, especially for chemicals with the least reactive modes of action; highly reactive chemicals are more problematic. In addition, QSAR or SAR models can serve as a decision-making framework tool to evaluate adequacy of data and are useful in cases where data availability is incomplete on the test and/or when there are differences in test methods. QSAR predictions and trend analysis of members in chemicals categories often use physico-chemical properties and/or other molecular descriptors relating to bioavailability and or reactivity to predict response variables of endpoints of interest. Application of such non test methods can be effective to minimise fish testing subject to validation and appropriate use within chemical domains of relevance.

178. The OECD QSAR Application Toolbox (available for download at http://www.oecd.org/document/54/0,3343,en_2649_34379_42923638_1_1_1_1,00.html) is a recently-developed and continuously evolving collection of databases and models aimed at "…filling gaps in (eco)toxicity data needed for assessing the hazards of chemicals." Features of this application include the identification of relevant structural characteristics and potential mechanism or mode of action of a target chemical, identification of other chemicals that have the same structural characteristics and/or mechanism or mode of action, and use of existing experimental data to fill the data gap(s).

179. Another valuable resource is the European Joint Research Centre (JRC) QSAR models database. This database is an inventory of information on the valid (Q)SAR models that have been submitted to the JRC, the intent of which is to provide an overview of available validated (Q)SAR models. The database provides a QSAR Model Reporting Format (QMRF), which is a harmonised template for summarising and reporting key information on (Q)SAR models, including the results of any validation studies. The information is structured according to the OECD principles for the validation of (Q)SAR models (http://ihcp.jrc.ec.europa.eu/facilities/JRC_QSAR_Model_Database.htm).

180. A particularly well-developed area of modelling is aquatic acute toxicity. Several different types of models exist for various classes of chemicals; a comprehensive review of QSAR strategies for aquatic toxicity is provided by Netzeva et al. (2007). See also "(Q)SAR Application Toolbox ver.1.1. "strategies for grouping of chemicals for data gap filling for acute aquatic toxicity endpoints", April 2010. An overview of the available aquatic toxicity data sources created to facilitate the development of QSAR models as well as a discussion of several QSAR expert systems is provided in Bradbury et al. (2003). However, as with use of any model, and particularly with QSAR models, care must be taken to use the model within its applicability domain, both with regard to chemical class as well as scientific purpose (Walker et al. 2003). Well-developed frameworks for modelling environmental fate and ecotoxicity related endpoints are embedded in the USEPA EPISUITE (USEPA 2008, at http://www.epa.gov/opptintr/exposure/pubs/episuite.htm) including acute and chronic aquatic toxicity to fish in the ECOSAR module.

181. Additional QSAR methods have been developed for acute fish toxicity testing, such as those developed by the UK DEFRA Alternatives to Animal Testing for Chemical Risk Assessment Project to assess narcotic modes of action (www.inchemicotox.org) and the development of global models including a fish acute toxicity model developed by the Danish QSAR group (cf. website with Danish QSAR model predictions: http://www.qsar.food.dtu.dk, which allow predictions with applicability domain indications on approximately 50.000 discrete organic EINECS chemicals). It should be pointed out that many QSAR models including those on acute fish toxicity generally apply only to a narrow chemical space (i.e. only have a limited Applicability Domain) and that such models provide reliable predictions when addressing well described and documented endpoints for structurally homologous series of compounds.

182. This guidance document focuses on fish toxicity testing and does not in detail address bioaccumulation testing strategies even though fish are the most widely used group of organisms employed in bioaccumulation testing. Such testing strategies, which also aim to reduce as much as possible the use of fish without compromising the purpose of obtaining an adequate basis for bioaccumulation assessment, have been published by ECHA (2008c) and in the scientific literature (e.g. Nendza and Müller (2010), Nendza and Herbst (2011)).

183. The US EPA has developed a SAR/decision tree for identifying chemicals within certain chemical classes (primarily food-use inerts and antimicrobial pesticides thus far) that bind to the

trout estrogen receptor alpha. This decision tree has now been favourably reviewed by the OECD and an EPA Scientific Advisory Panel (US EPA 2009b). Although the model is built using the trout receptor, there is quite a lot of supporting evidence that the trout and human cell systems behave identically − meaning this model can be used for both eco and human health applications. The EPA is now working on "coding" this model for computer automation and inclusion in the OECD toolbox (P. Schmieder, pers. comm.). Similar QSAR models have been developed for mammalian estrogen receptor binding and predictions of such models have, with applicability domain indications, been included into the OECD QSAR Application Toolbox and the Danish QSAR database. The same goes for an androgen reporter gene activation QSAR model.

5.8 *In Vitro*/*ex vivo* assays/high-throughput methods

184. There are numerous on-going efforts directed towards the development of novel assays to either predict or inform fish toxicity. These include *in vitro, ex vivo*, and various high-throughput methods.

5.8.1 Cell assays

185. *In vitro* fish cell assays from various tissues (both primary cells and immortalized cell lines) have been assessed by several groups (reviewed in Schirmer 2006, Castaño et al. 2003, as well as ECETOC 2005) as potential future alternatives to range finding testing in traditional *in vivo* tests. These assays facilitate the examination of toxicity pathways at the molecular and cellular levels, allow for tests on additional fish species, provide higher throughput, and the use of smaller volumes of test chemicals. They also serve to eliminate (cell lines) or reduce (primary cells) animal use.

186. Most assays utilizing fish cells have focused on acute lethality (Schirmer et al. 2006, Kramer et al. 2009), metabolism (see section below), and as a tool to better understand mechanisms of toxicity (Castaño et al. 2003). This latter use will help to elucidate adverse outcome pathways, which can help to create chemicals categories or QSAR models or to focus toxicity testing strategies (Ankley et al. 2010).

187. Much attention has been given to the development of predictive tools such as *in vitro* tests and *in silico* models and to alternatives to certain *in vivo* tests to detect endocrine modulation and disruption. The OECD's Validation Management Group for Ecotoxicity Testing (VMG-eco) recently developed a detailed review paper on receptor binding and transactivation assays in fish (OECD 2009). This review covers the use of both estrogen and androgen receptor assays in 14 different fish species. In addition, *ex vivo* assays such as liver slices (Schmieder et al. 2004) have been developed as useful tools to interpret the results of receptor binding assays, providing a high level of biological complexity while reducing animal use. A recently completed literature review on alternative endocrine disrupting chemicals tests in fish and amphibians, including an assessment of toxicogenomic approaches, provides a current assessment of the state of the science in this area [Scholz *et al* (2011)].

5.8.2 Embryo assays

188. Fish embryos are a desirable model for ecotoxicity testing due to the fact that they represent a complex biological system that can be assessed in a high-throughput manner (Scholz et al. 2008). An important aspect for fish embryo-based methods is the definition of protected and non-protected life stages of fish and, by that definition, whether such a method would be considered a replacement or refinement method in the strict sense of legislation. A review of the regulatory aspects regarding the use of fish embryos in environmental toxicology in various countries is

provided in Halder et al. (2010). The eleutheroembryo stage is post-hatch, but before the embryo is capable of independently feeding on exogenous food supplies and is a stage of on-going embryonic development. In some regulatory jurisdictions, the eleutheroembryonic period is regarded as a non protected life stage in this context.

189. The fish embryo toxicity (FET) test has been proposed as an alternative to the traditional fish acute toxicity test (OECD TG 203), and the lead country Germany has developed a draft OECD test guideline (OECD 2006). In 2005, the German Federal Environment Agency submitted the draft TG on the "Fish embryo toxicity (FET) test" to the OECD Test Guideline Program and a supportive Background Paper. Subsequently, OECD established the *ad hoc* Expert Group on the Fish Embryo Toxicity Test. Based on the outcome of expert meetings, OECD decided to perform a validation study (for details, see below). One concern is that the chorion can serve as a barrier that does not allow penetration of some test chemicals and that the test therefore might underestimate toxicity relative to that for juvenile or adult fish. Another concern is that the fish embryo is not sufficiently developed to possess a full spectrum of metabolism analogous to more developed fish life stages, and therefore results may for that reason differ between the FET and results from tests on juvenile or adult fish for chemicals which are metabolised by fish and where the metabolites have a significantly different toxicity profile than the parent compound. A recently-published review on the FET outlines data in this test for in a large number of compounds, comparing it to available *in vivo* data (Lammer et al. 2009).

190. Although the FET is currently developed for estimation of acute juvenile/adult fish toxicity, there is hope that assessment of additional endpoints and use of new technologies (such as "omics") in the future might allow assessment of chronic toxicity (Scholz et al. 2008, Voelker et al. 2007) and also that it may provide additional information related to chemical modes of action.

191. The OECD validation study, coordinated by the European Centre for the Validation of Alternative Methods (ECVAM), aims to evaluate the transferability as well as the inter- and intra-laboratory reproducibility of the test. Newly fertilised zebrafish eggs (20/concentration and control) are exposed for up to 96h (spanning the embryo and eleutheroembryo phases, but ending before the onset of exogenous feeding) to chemicals. Four apical endpoints are recorded daily as indicators of acute lethality in fish: coagulation of the embryo, lack of somite formation, non-detachment of the tail bud from the yolk sac and lack of heart-beat. LC_{50} values are calculated for 48h and 96h exposure. The report of Phase 1 of the validation has been published by OECD (OECD, 2011).

5.8.3 In vitro assays for bioaccumulation

192. *In vitro* assays to assess bioaccumulation in fish have been developed for both metabolism and uptake, which are the two main physiological processes that drive bioaccumulation. Chemical uptake in fish occurs through intestine, gills and skin. A recently established rainbow trout intestinal epithelial cell line (RTgutGC) method is an *in vitro* system to assess uptake of chemicals across the gut epithelium in fish (Kawano et al. 2010). Isolated and perfused fish intestine (Kleinow et al. 1998, Doi et al. 2000) and gills (Barron et al. 1989, Sijm et al. 1993) are *ex vivo* methods that can be used to estimate uptake *in vivo*; however, these tissues require a high degree of technical expertise to isolate and are only viable for a short time.

193. Several methodologies to determine metabolism have been developed to refine bioaccumulation assessments, because many of the traditional log K_{OW}-based BCF models currently in use are based on training sets of poorly metabolised substances (and so may over-predict BCFs for chemicals that are metabolised more extensively) (Nichols et al. 2007). These assays borrow from mammalian ADME (Absorption, Distribution, Metabolism, and Excretion) techniques used for

pharmaceutical development, and most focus on the liver as the main site of metabolism. Primary hepatocyte isolations from rainbow trout and carp have been used to predict *in vivo* metabolism to refine bioaccumulation estimates (Han et al. 2007, Cowan-Ellsberry et al. 2008), as have subcellular fractions such as fish S9 or microsomes. There are also several fish liver cell lines that demonstrate metabolic capacity (reviewed in Castaño et al. 2003), though these rates tend to be lower than those seen in primary liver cells or subcellular fractions (Weisbrod et al. 2009).

5.8.4 "Omics" technologies

194. Numerous studies utilizing "omics" (genomics, proteomics, transcriptomics, etc.) technologies have been studied in fish, and several reviews relating to the state of the science and potential use in regulatory ecotoxicology have recently been published (Ankley et al. 2006, 2008, Van Aggelen et al. 2010). Though the field is rapidly advancing, translation and interpretation of the data obtained in these studies is complex. These approaches do allow for increased knowledge concerning chemical modes of action and adverse outcome pathways, which can help to determine the toxicity tests that are necessary and appropriate for regulatory purposes. In addition, these techniques provide the ability to assess chemical effects at low doses.

195. The US EPA's National Centre for Computational Toxicology (NCCT) houses several programs whose focus is the development and implementation of high-throughput screens (HTS) for toxicological endpoints (US EPA 2010). NCCT itself has the capacity to run hundreds of cell-based assays using thousands of chemical preparations within days or weeks. The ToxCast^{TM} program is a collaboration between several labs providing a broad spectrum of cell, genomics or animal-based assays whose goal is to identify toxicity-related cellular pathways. Phase I of ToxCast screened 320 chemicals (309 different structures) in 467 different assays; interpretation of this data is on-going (Judson, et al. 2010). Although NCCT assays are based primarily on cells, proteins and genes from human and other mammals, ToxCast includes zebrafish and nematode assays. In addition, the technology and principles involved in data collection and interpretation would be transferable to assays developed from components derived from other species. Currently, the information from HTS and other NCCT screens are being used to prioritize chemicals for further testing and to identify the most likely major toxicity of chemicals. When pathway information becomes more complete, it will be possible to use the information to reduce uncertainties in data extrapolation and to inform risk assessment. NCCT is also developing two major publicly available, searchable databases, ACToR and ToxRefDB (http://actor.epa.gov/toxrefdb/faces/Home.jsp) to compile toxicological data from a variety of sources and includes ecotoxicological information.

196. The OECD, in a close cooperation with the International Programme on Chemical Safety (IPCS), has been working in the field of toxicogenomics and molecular screening with a focus on defining the needs and possibilities for the application of these emerging technologies in a regulatory context. The collaborative work between OECD and IPCS started by organizing twin workshops related to omics techniques for toxicology and eco-toxicology. The first workshop on human health aspects was held in November 2003 in Berlin with the IPCS as the leading organisation. The OECD took the lead in the organisation of the second workshop held in October 2004 in Kyoto, Japan, that focused on eco-toxicological aspects (OECD 2005). After this workshop, a follow-up survey was conducted on current toxicogenomic approaches available in member countries（OECD 2008).

197. In 2007, the OECD started the "Molecular Screening for Characterizing Individual Chemicals and Chemical Categories Project" (Molecular Screening Project). This project evaluates a number of selected chemicals in a series of molecular screening *in vitro* assays (High Throughput Screening (HTS)) with the aim of establishing a strategy for rationally and economically prioritizing chemicals for further evaluation based on molecular properties and categories linked to potential

toxicity. The project is led by the United States and supervised by the extended Advisory Group on Molecular Screening and Toxicogenomics. The US ToxCast^{TM} program (see below) forms a core part of the project. In 2008 and 2009 several subgroups related to specific pathways, mechanisms and effects as well as nomination of target chemicals and database development were set up under a cooperation between the US and other members.

5.9 Conclusions and recommendations

198. Several recommendations should be considered when developing new test guidelines, revising existing guidelines, and developing a tiered testing strategy for fish. These include the following:

- There is a strong need for consistency in the definitions of fish life-stages amongst test guidelines. These include, but are not limited to the definitions of embryo, eleutheroembryo, larva(e), juvenile, and adult stages.

- There are several well-developed and accepted methodologies that could be considered and employed as appropriate in a fish testing strategy concerning acute toxicity, such as:

  - use of the limit test (see chapter 5.4.2);

  - use of the threshold approach (see chapter 5.4.5);

  - use of screening methodologies that do not utilize animals (such as QSAR tools, chemicals categorisation, *in vitro* assays, the FET, or read-across)

Incorporation of "omics" technologies within existing *in vivo* tests can provide additional information regarding mechanism of action and adverse effect outcome (AoP), for example, genetic or proteomic information associated with particular adverse phenotypes. Also this may be used in development and validation of non animal methods.

- During the development of new, or revisions of existing, *in vivo* fish test guidelines, reduction, refinement and potentially replacement of animal use must be considered (see chapter 5.3). However, such changes should only occur if it is shown that they will not compromise the statistical power of the test.

- The number of test concentrations required by the test guideline should be optimised and/or the test design improved (e.g. by considering the EC_x (regression based) approach relative to the NOEC (hypothesis based) approach).

- In terms of the 3Rs, options for an optimised test design should be offered in all applicable cases.

Additional research on the development of fish alternatives is needed. Test strategies should be tiered when possible in order to best utilize resources and avoid unnecessary testing. Additional tests should not be required, if the outcome can be estimated from existing available data.

- Make more use of existing or confidential data to develop predictive models.

5.10 References

Ankley, G.T., Daston, G.P., Degitz, S.J., Denslow, N.D., Hoke, R.A., Kennedy, S.W., Miracle, A.L., Perkins, E.J., Snape, J., Tillitt, D.E., Tyler, C.R., Versteeg, D. (2006) Toxicogenomics in regulatory Ecotoxicology. Environ. Sci. Technol. 40: 4055-4065.

Ankley, G., Miracle, A., Perkins, E., Daston, D. (2008) Genomics in regulatory ecotoxicology. Pensacola, FL. SETAC Press.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mount, D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serrrano, J.A., Tietge, J.E., Villeneuve, D.L. (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem. 29: 730-741.

Arnot, J.A., Meylan, W., Tunkel, J., Howard, P.H., Mackay, D., Bonnell, M., Boethling, B.S. (2009) A quantitative structure-activity relationship for predicting metabolic biotransformation rates for organic chemicals in fish. Environ. Toxicol. Chem. 28: 1168-1177.

Asfaw, A., Ellersieck, M. R.,Mayer, F. L. (2003) Interspecies correlation estimations (ICE) for acute toxicity to aquatic organisms and wildlife. II. User manual and software. US Environmental Protection Agency: Washington, DC; EPA/600/R-03/106.

Barron, M.G., Schultz, I.R., Hayton, W.L. (1989) Presystemic branchial metabolism limits di-2-ethylhexulphalate accumulation in fish. Toxicol. Appl. Pharmacol. 98: 49-57.

Belanger, S.E., Dyer, S.D., Versteeg, D.J., Brill, J.L., Chaney, J.G., Raimondo, S J., Barron, M.G. (2009) Integrating non-vertebrate toxicological information as a substitute to vertebrate environmental testing. Seventh World Congress on Alternatives and Animal use in the Life Sciences, September, 2009, Rome, Italy.

Bradbury, S.P., Feijtel, T.C.J., van Leeuwen, C.J. (2004) Meeting the scientific needs of ecological risk assessment in a regulatory context. Environ. Sci. Technol. 38, 463A-470A

Castaño, A., Bols, N., Braunbeck, T., Dierickx, P., Halder, M., Isomaa, B., Kawahara, K., Lee, L.E.J., Mothersill, C., Pärt, P., Repetto, G., Riego Sintes, J., Rufli, H., Smith, R., Wood, C., Segner, H. (2003) The use of fish cells in ecotoxicology. ATLA 31: 317-351.

Cowan-Ellsberry, C.E., Dyer, S.D., Erhardt, S., Bernhard, M.J., Roe, A.L., Dowty, M.E., Weisbrod, A.V. (2008) Approach for extrapolating *in vitro* metabolism data to refine bioconcentration factor estimates. Chemosphere 70: 1804-1817.

de Wolf, W., Comber, M., Douben, P., Gimeno, S., Holt, M., Léonard, M., Lillicrap, A., Sijm, D., van Egmond, R., Weisbrod, A., Whale, G. (2007) Animal-use replacement, reduction, and refinement: development of an integrated testing strategy for bioconcentration of chemicals in fish. Integr. Environ. Assess. Manag. 3:3-17.

Doi, A.M., Lou, Z., Holmes, E., Li, C., Venugopalan, C.S., James, M.O., Kleinow, K.M. (2000) Effect of micelle fatty acid composition and 3,4,3',4'-tetrachlorobiphenyl (TCB) exposure on intestinal [^{14}C]-TCB bioavailability and biotransformation in channel catfish in situ preparations. Toxicol. Sci. 55: 85-96.

Dyer, S.D., Versteeg, D.J., Belanger, S.E., Chaney, J.G., Mayer F. L. (2006) Interspecies correlation estimates (ICE) predict protective environmental concentrations. Environ. Sci. Technol. 40: 3102-3111.

Dyer, S.D., Versteeg, D.J., Belanger, S.E., Chaney, J.G., Raimondo, S., Barron, M.G. (2008) Comparison of species sensitivity distributions derived from interspecies correlation models to distributions used to derive water quality criteria. Environ. Sci. Technol. 42: 3076-3083.

ECETOC (2005) Alternative testing approaches in environmental safety assessment. European Centre for Ecotoxicology and Toxicology of Chemicals, Techn. Report No. 97, 145 pp.

ECETOC (2007) Intelligent testing strategies in ecotoxicology: mode of action approach for specifically acting chemicals. European Centre for Ecotoxicology and Toxicology of Chemicals Techn. Report No. 102, 182 pp.

ECHA (2008a) Guidance on information requirements and chemical safety assessment Chapter R.7b: Endpoint specific guidance, and subchapters R.7a (human health endpoints), R.7b and R.7c (environmental endpoints). European Chemicals Agency Guidance for the implementation of REACH; available at: http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r7c_en.pdf?vers=20_08_08.

ECHA (2008b) Guidance on information requirements and chemical safety assessment Chapters R2 – R7: Information requirements. European Chemicals Agency Guidance for the implementation of REACH; available at: http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_en.htm?time=1282555491.

ECHA (2008c). Technical Guidance Document to Industry on Information Requirement for REACH: Chapter 7.10 Bioconcentration and bioaccumulation, p, 247-315

ECHA (2010) Practical Guide 10: How to avoid unnecessary testing on animals. ECHA-10-B-17-EN. European Chemicals Agency; ailable at:

http://echa.europa.eu/doc/publications/practical_guides/pg_10_avoid_animal_testing_en.pdf .

EU (1986) Directive 86/609/EEC of 24 November 1986 on the approximation of laws, regulations and administrative provisions of the Member States regarding the protection of animals used for experimental and other scientific purposes. Official Journal of the European Communities L358, pp. 1-29.

EU (2003) Council Directive 2003/15/EC of 27 February 2003 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products. Official Journal of the European Communities, L 66, 26. http://ec.europa.eu/enterprise/cosmetics/doc/200315/200315_en.pdf

EU (2006) Regulation 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Official Journal of the European Union L 136, pp. 3-280.

EU (2009) Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. Official Journal of the European Union, L 342, pp. 59- 209.

EU (2010) Directive 2010/63/EU of the European Parliament and the council of 22 September 2010 on the protection of animals used for scientific purposes. Official Journal of the European

Communities, L 275/33 (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:276:0033:0079:EN:PDF)

Grindon, C., Combes, R., Cronin, M., Roberts, D.W., Garrod, J.F. (2008) Integrated testing strategies for use with respect to the requirements of the EU REACH legislation. ATLA 36, Suppl. 1: 7–27.

Halder M, Léonard M, Iguchi T, Oris JT, Ryder K, Belanger SE, Braunbeck TA, Embry MR, Whale G, Norberg-King T, Lillicrap A (2010), Regulatory aspects on the use of fish embryos in environmental toxicology. Integr Environ Assess Manag. 2010 Jul;6 (3):484-91

Han, X., Nabb, D.L., Mingoia, R.T, Yang, C.H. (2007) Determination of xenobiotic intrinsic clearance in freshly isolated hepatocytes from rainbow trout and rat and its application in bioaccumulation assessment. Environ. Sci. Technol. 41: 3269-3276.

Hutchinson, T.H., Barrett, S., Buzby, M., Constable, D., Hartmann, A., Hayes, E., Huggett, D., Länge, R., Lillicrap, A.D., Straub, J.O., Thompson, R.S. (2003) A strategy to reduce the numbers of fish used in acute ecotoxicity testing of pharmaceuticals. Environ. Toxicol. Chem. 22: 3031-3036.

Hutchinson, T. (2008) Intelligent testing strategies in ecotoxicology: approaches to reduce and replace fish and amphibians in toxicity testing. NC3Rs #14. Available at www.nc3Rs.org.uk.

Jaworska, J., Gabbert, S., Aldenberg, T. (2010) Towards optimisation of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies. Regul. Toxicol. Pharmacol. 57:157-67.

Jeram, S., Sintes, J.M., Halder, M., Fentanes, J.B., Sokull-Klüttgen, B., Hutchinson, T.H. (2005) A strategy to reduce the use of fish in acute ecotoxicity testing of new chemical substances notified in the European Union. Regul. Toxicol. Pharmacol. 42: 218-24.

Judson, R.S., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Mortensen, H.M., Reif, D.M., Rotroff, D.M., Shah, I., Richard, A.M., Dix, D.J. (2010) *In vitro* screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Environ. Health Perspect. 118: 485-492.

Kawano, A., Haiduk, C., Schirmer, K., Hanner, R., Lee, L.E.J., Dixon, B., Bols, N.C. (2010) Development of a rainbow trout intestinal epithelial cell line and its response to lipopolysaccharide. *Aquac. Nutr*., in press (http://dx.doi.org/10.1111/j.1365-2095.2010.00757.x)

Kleinow, K.M., James, M.O., Tong, Z., Venugopalan, C.S. (1998) Bioavailability and biotransformation of benzo[a]pyrene in an isolated perfused in situ catfish intestinal preparation. Environ. Health Perspect. 106:155-166.

Kramer, N.I., Hermens, J.L., Schirmer, K. (2009) The influence of modes of action and physicochemical properties of chemicals on the correlation between *in vitro* and acute fish toxicity data. Toxicol. in Vitro. 23: 1372-1379.

Lammer, E., Carr, G.J., Wendler, K., Rawlings, J.M., Belanger, S.E., Braunbeck, T. (2009) Is the fish embryo test (FET) with the zebrafish (*Danio rerio*) a potential alternative for the fish acute toxicity test? *Comp Biochem Physiol* 149C: 196-209.

National Academy of Sciences (2007) Toxicity testing in the twenty-first century: a vision and a strategy. Committee on Toxicity and Assessment of Environmental Agents, National Research Council. ISBN: 0-309-10989-2, 146 pp.

Nendza, M. and Herbst, T. (2011). Screening for low aquatic bioaccumulation (2): physico-chemical constrains, SAR & QSAR Env. Res.., 22 (3-4), 351-64

Nendza, M. and Müller, M. (2010). Screening for low aquatic bioaccumulation (1) Lipinski´s "Rule of 5" molecular size, SAR & QSAR Env Res. 21, p 495-512

Netzeva, T., Pavan, M., Worth ,A. (2007) Review of data sources, QSARs, and Integrated Testing Strategies for aquatic toxicity. European Commission Joint Research Centre Scientific and Technical Report, EUR 22943 EN -2007.

Nichols, J., Erhardt, S., Dyer, S., James, M.O., Moore, M., Plotzke, K, Segner, H., Schultz, I., Vasiluk, L., Weisbroad, A. (2007) Workshop report: Use of *in vitro* absorption, distribution, metabolism, and excretion (ADME) data in bioaccumulation assessments for fish. Hum. Ecol. Risk Assess. 13: 1164-1191.

OECD (1994) US EPA/EC Joint Project on the evaluation of (Quantitative) Structure Activity Relationships. Environment Monograph No. 88, Paris, France 81p.

OECD (1998) Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems Test Test Guideline No 212: Fish, Short-term Toxicity Test on Embryo and Sac-Fry Stages. OECD, Paris.

OECD (2000), Guidance Document on the Recognition, Assessment, and use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation, Series on Testing and Assessment No. 19, ENV/JM/MONO(2000)7, OECD, Paris

OECD (2005), Report of the OECD/IPCS Workshop on Toxicogenomics, OECD Series on Testing and Assessment No.50, ENV/JM/MONO(2005)10. OECD, Paris.

OECD (2006) Draft Proposal for a New Guideline, Fish Embryo Toxicity (FET) Test. OECD Guideline for the Testing of Chemicals. Organisation for Economic Cooperation and Development, Paris, France.

OECD (2008), Report of the Second Survey on Available Omics Tools, OECD Series on Testing and Assessment No.100, ENV/JM/MONO(2008)35. OECD, Paris.

OECD (2010), Report of the Focus Session on Current and Forthcoming Approaches for Chemica Safety and Animal Welfare, Series on Testing and Assessment No. 113, ENV/JM/MONO(2010)5, OECD, Paris

OECD (2010) OECD Guidance document No 126: Short Guidance on the Threshold approach for Acute Fish Toxicity – ENV/JM/TG(2010)/7. http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=ENV/JM/MONO(2010)17&doclanguage=en. http://www.oecd.org/dataoecd/49/3/43226061.pdf.

OECD (2011). Validation Report (Phase 1) for the Zebrafish Embryo Toxicity Test, No.157, Series on Testing and Assessment - ENV/JM/MONO(2011)37

Raimondo, S., Vivian, D.N., Barron, M.G. (2010) Web-based Interspecies Correlation Estimation (Web-ICE) for Acute Toxicity: User Manual. Version 1.1. EPA/600/R-10/004. Gulf Breeze, FL; available at: http://www.epa.gov/ceampubl/fchain/webice/index.html.

Rufli, H. and Springer, T.A. (2011). Can we reduce the number of fish in the OECD acute toxicity test? Environ. Toxicol. Chem. 30: 1006-1011.

Rufli, H. (2012) Introduction of moribund category to OECD Fish Acute Test and its effect on suffering and LC50 values. Environ Toxicol Chem, 31(12), pp. 1-6

Russell, W.M.S., Burch, R.L. (1959) The principles of humane experimental technique. Methuen, London, 253 pp.

Schirmer, K. (2006) Proposal to improve vertebrate cell cultures to establish them as substitutes for the regulatory testing of chemicals and effluents using fish. Toxicology 224: 163-183.

Schmieder, P.K., Tapper, M.A., Denny, J.S., Kolanczyk, R.C., Sheedy, B.R., Henry, T.R., Veith, G.D. (2004) Use of trout liver slices to enhance mechanistic interpretation of estrogen receptor binding for cost-effective prioritization of chemicals within large inventories. Environ. Sci. Technol. 38: 6333-6342.

Scholz, S., Fischer, S., Gundel, U., Juster, E., Luckenbach, T., Voelker, D. (2008) The zebrafish embryo model in environmental risk assessment – application beyond acute toxicity testing. Environ. Sci. Pollut. Res. Int. 15: 394-404.

Scholz, S., Renner, P., Ortego, L.S., Belanger S., Busquet, F., Davi, R., Demeneix, B., Denny, J.S., Leonard, M., McMaster, M., Villeneuve, D., Embry, M. (2011) Alternatives to *in vivo* tests to detect endocrine disrupting chemicals (EDCs) in fish and amphibians. HESI Animal Alternatives in Environmental Risk Assessment Project Committee Commissioned Report. Submitted to Environmental Research in Toxicology.

Sijm, D., Part, P., Opperhuizen, A. (1993) The influence of temperature on the uptake rate constants of hydrophobic compounds determined by the isolated perfused gills of rainbow trout (*Oncorhynchus mykiss*). Aquat. Toxicol. 25: 1-14.

Springer, T., Guiney, P., Krueger, H., Jaber, M. (2008) Assessment of an approach to estimating aquatic bioconcentration factors using reduced sampling. Environ. Toxicol Chem. 27: 2271-2280.

Stephan, C.E. (2002) Use of species sensitivity distributions in the derivation of water quality criteria for aquatic life by the U.S. Environmental Protection Agency. In *Species Sensitivity Distributions in Ecotoxicology*, Lewis Publishers, Boca Raton, FL, pp. 211-220.

US EPA (2000) Voluntary Children's Chemical Evaluation Program (VCCEP), Office of Pollution Prevention and Toxics. http://www.epa.gov/oppt/vccep/pubs/basic.html#tiers.

US EPA (2008) EPISUITE. Version 4.0, Washington, DC, USA; available at: http://www.epa.gov/opptintr/exposure/pubs/episuite.htm.

US EPA (2009a) The US Environmental Protection Agency's Strategic Plan for Evaluating the Toxicity of Chemicals. Office of the Science Advisor, Science Policy Council, U.S. EPA, Washington, DC; http://www.epa.gov/spc/toxicitytesting/.

US EPA (2009b) An Effects-based Expert System to Predict Estrogen Receptor Binding Affinity for Food Use Inert Ingredients and Antimicrobial Pesticides: Application in a Prioritization Scheme for Endocrine Disruptor Screening. Office of Pesticide Programs, US EPA, *Washington, DC.* Available at Regulations.gov, Docket number: EPA-HQ-OPP-2009-0322, Document number: 002.

US EPA (2010) National Center for Computational Toxicology (NCCT). Office of Research and Development. http://www.epa.gov/ncct/ (accessed 7/1/2010).

US Senate (2010) US Senate bill S.3209 - Safe Chemicals Act of 2010; available at: http://thomas.loc.gov.

Van Aggelen, G., Ankley, G., Baldwin, W., Bearden, D.W,, Benson, W.H., Chipman, J.K, Collette, T.W., Craft, J.A., Denslow, N.D,, Embry, M.R., Falciani, F., George, S.G., Helbing, C.C., Hoekstra, P.F., Iguchi, T., Kagami, Y., Katsiadaki, I., Kille, P., Liu, L., Lord, P.G., McIntyre, T., O'Neill, A., Osachoff, H., Perkins, E.J., Santos, E.M., Skirrow, R.C., Snape, J.R., Tyler, C.R., Versteeg,D., Viant, M.R., Volz, D.C., Williams, T.D., Yu, L. (2010) Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. Environ. Health Perspect. 118: 1-5.

Van Leeuwen, C.J., Patlewicz, G.Y., Worth, A.P. (2007) Intelligent Testing Strategies. In: van Leeuwen, C.J., Vermiere, T.G. (eds.). Risk Assessment of Chemicals: An Introduction. Springer. pp. 467-509.

Voelker, D., Vess, C., Tillmann, M., Nagel, R., Otto, G.W., Geisler, R., Schirmer, K., Scholz, S. (2007) Differential gene expression as a toxicant-sensitive endpoint in zebrafish embryos and larvae. Aquat. Toxicol. 81: 355-364.

Walker, J. D., Jaworska, J.S., Comber, M.H.I., Schultz, T.W., Dearden, J.C. (2003) Guidelines for developing and using Quantitative Structure-Activity Relationships. Environ. Toxicol. Chem. 22: 1653-1656.

Weisbrod, A.V., Sahi, J., Segner, H., James, M.O., Nichols, J., Schultz I., Erhardt, S., Cowan-Ellsberry, C., Bonnell, M., Hoeger, B. (2009) The state of *in vitro* science for use in bioaccumulation assessments for fish. Environ. Toxicol Chem. 28: 86-96.

Weyers, A., Sokull-Klüttgen, B., Baraibar-Fentanes, J., Vollmer, G. (2000) Acute toxicity data: a comprehensive comparison of results of fish, *Daphnia* and algae tests with new substances notified in the EU. Environ. Toxicol. Chem. 19, 1931-1933. 6. Review of existing OECD guidelines and guidelines in preparation

6. REVIEW OF EXISTING OECD TEST GUIDELINES AND GUIDELINES IN PREPARATION

6.1 Introduction

199. This chapter contains reviews of all existing OECD Test Guidelines (TGs) in the order they are numbered in the published Test Guideline series. In addition, reviews of proposed test guidelines (latest drafts available at the time of writing) in the OECD programme are included at the end of the chapter.

The reviews are presented in tabular form, and include sections on:

- *Deliverables* (*what data/information does the TG deliver (acute/chronic, endpoints, etc.?))*
- *Prerequisites (data listed as required to conduct the test)*
- *Strengths of the Guideline*
- *Limitations of the Guideline (What are the limits of the data/information?)*
- *Statistics (Is the guidance presented in the TG current and in line with OECD 2006? Consideration of the benefits/practicalities of square root allocation)*
- *Terminology* (*Is all terminology current, are the TGs consistent e.g. describing life stages etc.?)*
- *Concentration setting (Is there sufficient guidance on how to choose test concentrations? Are the TGs consistent in respect to limit concentrations? Should guidance on potential range finding strategies be elucidated?)*
- *Quality Assurance (Which criteria are required as validity criteria?)*
- *Animal Minimisation (Do the TGs sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?)*
- *Non-solvent delivery/solvent use* (*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?)*
- *Species effectiveness* (*Is it justified to have all species as potential test organisms? When are tests on multiple species required?)*
- *General (Are there any points of clarity/interpretation required following experience of the TGs in regulatory application?)*

6.1.1 General Statistical Considerations

200. The test guidelines that consider toxicity to fish species (i.e., all except OECD TG 305) are all designed to measure a No Observed Effect Concentration (NOEC), a Lowest Observed Effect Concentration (LOEC), or some value, x, of an Effect Concentration (EC_x). There are general statistical considerations that must be made for these test designs, but which are not always described in the TGs. These considerations are described below.

6.1.2 Design for NOEC/LOEC or EC_x

201. The replicate or unit of analysis is the tank/test vessel, unless otherwise justified. Specify the size effect to be determined from the experiment. For an NOEC/LOEC, adjust the design so that it has 75 - 80 % power to detect the specified magnitude effect. The procedure for this is described in Chapter 3 of this document. For an EC_x (e.g. OECD TG 203), adjust the design so that the effect size, x, is not in the confidence interval for the control mean response. The procedure for this is described in Chapter 3 of this document.

202. If a solvent is used, then refer to the appropriate regulatory authority concerning the choice of control to be used. For effect size, x, use scientific judgement as to whether a monotone concentration-response is expected. If no mortality is observed in the control, a standard model can be fitted. In case mortality is observed in the control, a parameter to estimate the background mortality should be included in the model (see OECD 2006). For an EC_x, OECD (2006) provides a method for determining the right model(s) to fit for this purpose. Chapter 3 provides additional details for determining whether a particular value of x is appropriate for the design. For a NOEC, consult Chapter 3 for the appropriate statistical test to use.

6.2 OECD TG 203: Fish, acute toxicity test (adopted 17th July 1992)

Category	Description
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* Acute exposure assessing impacts on mortality. • acute toxicity of substance to fish after 24, 48, 72 and 96 h (determination of LC_{50} value) • maximum concentration causing no (LC_0) and 100 % mortality (LC_{100}; NOT used in risk assessment; may require additional test concentration)
Prerequisites	Information on the test substance: • physicochemical data of test substance including water solubility, stability and biodegradability (OECD TG 301) • reliable analytical method of chemical analysis of test concentrations
Strengths	• second major update of a guideline originally adopted in 1981 and first updated in 1984 • data available for almost any substance ever tested in aquatic toxicology.
Spectrum of test substances	• generally, no restriction • TG can be applied to any of the substance types as described in Chapter II assuming the physicochemical properties of the test item allow for testing
Limitations	*What are the limits of the data/information?* • no clear statement as to the preference of flow-through conditions; vague wording ("constant conditions *should* be maintained") • no positive control (for the sake of animal number reduction?) • test species recommended only cover freshwater species • multiple species testing mentioned without guidance as to when this should be performed • no guidance about the use of fish previously treated against disease or time required after prophylactic treatment (if disease does not induce mortality; § 23) • discussion of mortality issues and holding arrangements over 48h prior to test initiation may be too short • no restriction with respect to feeding regime (no quality control of fish food used) • no guidance on range finding ("range-finding test properly conducted…", § 17) • maximum fish load for flow-through conditions not defined ("may be higher than 1.0 g fish/litre") • preference for test duration of 96 hours; no guidance as to under which conditions another test duration should be used • no mention of test tank replication in TG • lack of guidance on analysis of dilution water control and solvent control • no recommendation as to how effects by solvents should be considered (only reporting of "incidents in the course of the test which might have influenced the results", § 23)

Category	Description
Statistics	*Is the guidance presented in the TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • no recommendation as to which statistical methods should be used ("normal statistical procedures…", § 21) • no indication of the unit of comparison (individual fish or replicate test vessel) is made • no guidance as to when data obtained should be classified as "inadequate for the use of standard methods of calculating the LC_{50})" (§22)
Terminology	*Is all terminology current, is the TG consistent e.g. describing life stages etc.* • scientific name for zebrafish has changed from *Brachydanio rerio* to *Danio rerio* • most common name of *Oryzias latipes* is "Japanese medaka" (rather than "ricefish") • no definition as to what is meant by "fish should be in good health"; no guidance about use of fish previously treated against disease • no clear definition of "*good quality* natural water" (§ 12) • fish length terminology is not defined (e.g. standard, fork or total) • no guidance as to when data obtained should be classified as "*inadequate* for the use of standard methods of calculating the LC50)" (§22)
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • limit test concentration: 100 mg/L (full test, if any mortality in the limit test) • at least 5 concentrations in a geometrical series with a factor preferably not exceeding 2.2 • no guidance on range finding given ("range-finding test properly conducted…", § 17)
Quality assurance	*Which criteria are required as validity criteria?* • mortality in controls should not exceed 10 % (or 1 individual, if less than 10 fish are used per concentration) • no positive control (for the sake of animal number reduction?) • measured concentrations should be within ± 20 % deviation from nominal concentrations; if > 20 % deviation, results should be given with reference to measured concentrations • dissolved oxygen ≥ 60 % • overall mortality over 7 d acclimatization prior to test initiation ≤ 10 % → rejection of entire batch • overall mortality over 7 d acclimatization prior to test initiation between 5 and 10 % → prolongation of acclimatization to 14 d
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • if compared to older versions, significant reduction in the number of animals used per concentration down to 7 individuals • further reduction by extension of the concentration range by allowing a spacing factor of 2.2 instead of 2

Category	Description
	• introduction of a limit test at 100 mg/L of test substance • no positive control • no guidance as to how use records of abnormalities (e.g. loss of equilibrium, swimming behaviour, respiratory function, pigmentation…, § 19) • no mention of test tank replication • determination of maximum concentration causing no (LC_0) and 100 % mortality (LC_{100}) require additional test concentration; since this information is not used in risk assessment, there is animal use and ethics concerns • since LC_0 and LC_{100} are not used in risk and hazard assessment as well as for labelling, determination questionable (additional concentration, if outside range tested) • as a possible consideration for future revision, the number of fish per concentration could be decreased to 6 fish per concentration without loss of precision (further details in Rufli & Springer, 2011) • As a general recommendation, better guidance for range-finding is required • Fish Embryo Test (FET) might at least serve for range-finding • currently, the Fish Embryo Test (FET) is under validation for use as alternative to TG 203
Non-solvent delivery / Solvent use	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • no clear statement as to preference for non-solvent delivery • no recommendation as to how effects by solvents should be considered • only ultrasonic dispersion as alternative preparation techniques mentioned
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • basically no restriction in the choice of test species (§ 8) • however, table at the end of TG only recommends freshwater species as test organisms • multiple species testing mentioned without guidance as to when this should be performed • several studies show that differences in the intrinsic toxicity of a substance between species are negligible and that there is no reason to use other species than the recommended species in the TG in order to find the intrinsic toxicity to fish; no need to use exotic species
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* • Separate measurement of LC_{100} as a deliverable should be removed for ethical concerns (however, higher and lower concentrations than LC_{50} needed for accurate interpolation of LC_{50}) • Harmonization with USEPA 850.1075 desirable • Weight criteria be added to the OECD guideline (as USEPA 850.1075), but length could also be referenced • Indication as to the unit of comparison (individual fish or replicate test vessel)

Category	Description
	• Clarification whether mortality as an endpoint requires replicates • As a general consideration for TGs with mortality endpoint: definition of "mortality" and "moribundity" and clarification as to which parameter/term should be used • General recommendation to consider harmonising guidance with respect to solvents and dispersants ; in case of solvent use, actual amount should be minimised as far as practically possible (instead only of definition of maximum solvent, emulsifier or dispersant at a maximum concentration of 100 mg/L) • Addition of language with respect to test extension beyond 96h if mortality slowly increasing, approaching LC50 (however, feeding might become an issue and raise potential ethical committee problems)

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

Category	Description
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc)?* Prolonged (≥ 14 days) exposure assessing impacts on mortality. • threshold levels of lethal and other observed effects • if required, extension of test period by one or two weeks • mortality over ≥ 14 days • NOEC for lethal and non-lethal effects • effects other than lethal ones: all effects observed on appearance, size and behaviour ("clearly distinguishable from the control animals, e.g. swimming behaviour, reaction to external stimuli, changes in appearance of fish, reduction or cessation of food intake, changes in length or body weight")
Prerequisites	• Physicochemical data of test substance
Strengths	• If compared to OECD TG 203 (Fish, acute toxicity test), consideration of sublethal effects in addition to mortality over ≥ 14 days.
Spectrum of test substances	• generally, no restriction • TG can be applied to any of the substance types as described in Chapter 2 assuming the physicochemical properties of the test item allow for long-term testing
Limitations	*What are the limits of the data/information?* In many aspects, the wording of TG 204 appears unspecific and lacks precision: • test does not specify life-stages, e.g. phases of sexual differentiation or reproduction, which might be important for the assessment of specific effects such as endocrine disruption • no clear statement as to the preference of flow-through conditions; vague wording ("there *should* be evidence that the concentration of the substance being tested has been satisfactorily maintained") • no guidance for intervals of test substance renewal in semi-static procedures • no positive control (reference compound) • limitation to freshwater species; test species recommended only cover freshwater species (→ OECD TG 203) • multiple species testing mentioned without guidance as to when this should be performed • no guidance about the use of fish previously treated against disease ("should be avoided, but reported when used") • no information about number and spacing of test concentrations • no guidance as to which "appropriate procedures other than analysis for giving evidence that adequate concentrations of the test substance have been maintained" might be appropriate • no restriction with respect to feeding regime (no quality control of fish food used) • no guidance (or even mention) on selection of concentrations or range finding • maximum fish loading for flow-through conditions not defined ("may

Category	Description
	be higher than 1.0 g fish/litre") • preference for test duration of 14 days; no guidance as to under which conditions test duration should be extended • no mention of test tank replication • lack of guidance whether the unit of statistical comparison should be individual fish or replicate test tank • lack of guidance on analysis of dilution water control and solvent control • imprecise wording: "*representative samples* of test population should be weighed and measured before test start" • no recommendation as to how effects by solvents should be considered (only reporting) • no guidance as to which measures should be taken, if oxygen saturation drops to < 60 %
Statistics	*Is the guidance presented in the TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • no guidance for statistical evaluation of data • no indication of the unit of comparison (individual fish or replicate test vessel) is made • lack of guidance whether the unit of statistical comparison should be individual fish or replicates of test tanks
Terminology	*Is all terminology current, is the TG consistent, e.g. describing life stages etc.?* • no definition as to what is meant by "fish should be in good health"; no guidance about use of fish previously treated against disease • no clear definition of "*good quality* natural water" (§ 12, OECD TG 203) • fish length terminology is not defined (e.g. standard, fork or total; cf. table OECD TG 203) • no precise recommendation for intervals of observations: "It is *desirable* that daily records be kept of all observed effects, but a minimum of three observation sessions per week must be conducted."
Concentration setting	*Is there sufficient guidance on how to choose test concentrations; is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated?* • limit test concentration: 100 mg/L (full test, if any mortality in the limit test) • no information about number and spacing of test concentrations • no guidance (or even mention) on selection of concentrations or range finding
Quality assurance	*Which criteria are required as validity criteria?* • mortality in controls should not exceed 10 % at the end of the test • no positive control (reference substance) • measured concentrations should be within ± 20 % deviation from nominal concentrations; if > 20 % deviation, results should be given with reference to measured concentrations • dissolved oxygen ≥ 60 % throughout test (semi-static procedures: aeration allowed, provided it does not lead to a significant loss of test substance)

Category	Description
	• dissolved oxygen ≥ 80 % during acclimatization • following 48 h settling-in period, overall mortality over 7 d acclimatization prior to test initiation ≤ 10 % → rejection of entire batch • overall mortality over 7 d acclimatization prior to test initiation between 5 and 10 % → prolongation of acclimatization to 14 d • flow-through → concentrations of the substance in test solutions *may* be determined (once?) at beginning of test • semi-static test → concentration to be verified at beginning, immediately prior to first renewal of test solution and at termination of test
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • at least 10 for each concentration and control • limit test concentration: 100 mg/L • no reference to potential minimization of number of test animals • no mention of test tank replication
Non-solvent delivery / Solvent use	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • no clear statement as to preference for non-solvent delivery • no recommendation as to how effects by solvents should be considered • only ultrasonic dispersion as alternative preparation techniques mentioned • no recommendation that if a solvent is used that the actual amount should be minimised as far as practically possible; only definition of maximum solvent/emulsifier/dispersant concentration of 100 mg/L
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • restriction to freshwater test species • reference to Table given in OECD TG 203 • OECD TG 203: table at the end of TG only recommends freshwater species as test organisms • use of non-standard OECD fish species (if there is evidence of significantly higher sensitivity) allowed (however, no or little guidance on potential modifications of test procedures, if required) • multiple species testing mentioned without guidance as to when this should be performed
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* • since there is routinely an option in acute testing guidelines (TG 204) to continue this observation period if mortalities continue to occur of an exposure period or as deemed necessary by the researcher, it does not seem necessary to maintain a separate acute toxicity guideline for the sole purpose of having a longer exposure period • TG 204 only very rarely used • recommendation to remove TG 204

6.4 OECD TG 210: Fish, early-life stage toxicity test (adopted 17th July 1992)

Category	Description
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* Chronic exposure assessing impacts (lethal and sub-lethal) on fish early life-stages (embryos, larvae and juveniles) on mortality, hatch, growth and development. • Hatch success (%) • Time to hatch (time; days) • Abnormal appearance; deformity (number) • Abnormal behaviour (qualitative) • Survival/ mortality at embryo, larval, juvenile stages and overall (%) • Weight (mass; group weights for small species) • Length; standard, fork or total • NOECs and LOECs (if possible) should be determined for each of the responses assessed assuming the data are amenable to statistical analysis
Strengths	• can be applied to any of the substance types as described in Chapter II assuming the physicochemical properties of the test item allow for long-term testing • robust established methodology that assesses impacts of the test item on specific life-stages that are known to be sensitive (covers many critical life events) • established relationship between the results of early life-stage tests and other relevant partial or full lifecycle tests (e.g., McKim 1977)
Limitations	*What are the limits of the data/information?* • guidance to thinning of larvae shortly after hatch not described: in practise, some laboratories initiate a test with an excess of embryos and indiscriminately thin to a set number after hatch (approach can be useful as it ensures an equal number of individuals per replicate progress to the larval-juvenile stage and potentially reduces inter-replicate variability) • For trout studies, it is not possible to determine if all the eggs initiating the exposure are fertilised (development not observable); therefore, it is not possible to determine if the batch of eggs was of sufficient quality or that sufficient individuals are available to meet the statistical requirements for all endpoints. One approach is to add a viability control to assess the percent fertilisation of the batch of eggs. This can be destructively sampled at ca 14-days by clearing the sample in acetic acid and looking for appearance of the neural keel to confirm fertilisation. Additional guidance for best practise in trout studies may be considered. • It is a TG requirement to record abnormal appearance. However, the TG lacks examples of commonly observed deformities (e.g. lordosis, scoliosis, kyphosis etc). • fish length terminology not defined (e.g. standard, fork or total lengths) • scientific name for zebrafish has changed from *Brachydanio rerio* to *Danio rerio* • incongruence between the use of developmental stages (e.g. before

Category	Description
	cleavage of blastodisc) and common practise of using time based descriptors (e.g. <24 hours old) • loading rates may require an additional caveat for a minimum tank size: initial analysis suggests that tanks sizes < 7 L may impact growth even if loading rates are met • loading (§ 18) suggests that the flow rate should be such that at least 60 % of the oxygen air saturation value be achieved without aeration; This may not be achievable depending on the properties of the test substance ─ additional wording required • TG sensitive to non-specific toxicants and some developmental toxicants, but not to some endocrine disrupters, since test does not cover sexual maturation and reproduction • since egg stage may be resistant to the absorption of some toxicants, test may not be sensitive to some substances that have the ability to interfere with embryonic development (to be explored further) • time to hatch is not well described in terms of how it should be reported
Statistics	*Is the guidance presented in the TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* Statistical guidance is not prescribed due to potential variations in test design (§ 33 and 34). • analysis of variance and contingency tables recommended, but specific guidance is lacking • in practise, certain regulatory agencies require monotonic data to be analysed by William's test – yet, method not mentioned in the statistical analysis section • no indication of the unit of comparison (individual fish or replicate test vessel) • no guidance on analysis of dilution water control and solvent control
Terminology	*Is all terminology current, is the TG consistent e.g. describing life-stages etc.* Life stage terminology • limited to embryos, larvae and juveniles • § 17: duration uses developmental stage (before cleavage of blastodisc), but typical usage is embryos <24 hours old Other • scientific name for zebrafish has since changed from *Brachydanio rerio* to *Danio rerio* • fish length terminology not defined (e.g. standard, fork or total)
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • TG states that a fish acute toxicity test (OECD TG 203) preferably in the test species, water solubility, vapour pressure and analytical method should be available • test concentrations determined by the need to achieve a NOEC (and preferably LOEC) • limit test concentration (§ 21): concentrations of the substance higher than the 96-hour LC_{50} or 10 mg/L, whichever is the lower, need not be tested

Category	Description
	• no guidance on range finding
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • minimum numbers of eggs per treatment level are given (§ 18; at least 60 eggs divided equally between 2 replicates/treatment • however, flexibility in the number of replicates/treatment in practise means that this minimum may not be seen as the optimal number; further guidance needed for optimising statistical power whilst minimising the number of fish used • § 28 ─ abnormal appearance ('Abnormal animals should only be removed from the test vessels on death'): as a requirement of some national animal welfare regulatory bodies (e.g. UK Home Office) under certain circumstances, deformity may be so severe that the animal should be removed before death and terminated to avoid suffering; therefore, guidance allowing for the minimisation of suffering and meeting local legislative requirements required
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • Preference for non-solvent delivery is clearly stated (§ 22). • § 6 validity criterion on solvent effect unclear as to what is considered adverse: '... nor produce any other adverse effects on the early-life stage as revealed by a solvent only control.' • No mention or reference to alternative preparation techniques. • No recommendation that if a solvent is used that the actual amount should be minimised as far as practically possible.
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • Species split into recommended (rainbow trout, fathead minnow, zebrafish, Japanese medaka and sheepshead minnow) and other well documented species (see table 1B). • tests with rainbow trout are generally longer to perform, seasonal in terms of embryo (or gamete) availability and vary in fertilisation success (which can only be confirmed well into the test); this should be considered opposite the advantages of using warmer water species (fathead minnow, zebrafish and Japanese medaka) • no evidence that a certain species is systematically more sensitive than the other commonly used species • very rare for tests on multiple species to be performed • however, potentially testing on a non-standard OECD species may be required if that species has been shown to be acutely significantly more sensitive (Annex 5 contains guidance for other species)
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* • loading rates may require an additional caveat for a minimum tank size: initial analysis suggests that tanks sizes < 7 L may impact growth even, if loading rates are met • § 18: loading suggests that the flow rate should be such that at least 60 % of the oxygen air saturation value is achieved without aeration; this may not be achievable depending on the properties of the test

Category	Description
	substance • test initiation should be based on stage rather than time (post-fertilisation) • Guidance on typical time for stages by species could be added • temperature effects on hatching need to be considered (e.g. long-range transport of eggs, especially rainbow trout, may be a problem in practice) • flexibility in the number of eggs per replicate may be indicated, in addition to the number of replicates, when considering the statistical power of the test • different validity criteria by species may be needed • more detail regarding range-finding should be added • basis for measurement of individuals should be reconsidered; wet weight may be better than dry weight

6.5 OECD TG 212: Fish, Short-term Toxicity Tests on Embryo and Sac-fry Stages (adopted 21st September 1998)

Category	Description
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* Sub-chronic exposure of fish embryos and larvae assessing mainly impacts on mortality, but also on growth and development. • hatch success (%) • time to hatch (time; days) • abnormal appearance; deformity (number) • abnormal behaviour (qualitative) • survival and mortality at embryo, larval, stages and overall (%) • length (and weight)
Strengths	• Pre-test for fish-early life stage test • The fish embryo and sac-fry test is a short and mostly reliable pre-test for the fish early life stage test and to some extend to the fish sexual development test. Also tests with multiple species are possible. • The TG can be applied to any of the substance types assuming the physicochemical properties of the test item allow for larval exposure (e.g. no oil film).
Limitations	*What are the limits of the data/information?* • it is expected that the embryo and sac-fry test would be less sensitive than the Full Early Life-Stage Test, particularly with respect to chemicals with a high lipophilicity (log P_{OW} > 4) and chemicals with a specific mode of toxic action • since egg stage may be resistant to the absorption of some toxicants, test may not be sensitive to some substances that have the ability to interfere with embryonic development • not many data for this test design exist, as it is mainly used as a pre-test • no guidance for data interpretation • better estimation of test item-related effects would be possible, if only fertilised eggs were used; in the case of trout it is not possible to distinguish fertilised eggs at the start of the test • during the embryonic development (at least for common warm-water species), assessment of sub-lethal effects in well plates should be considered (single eggs can be observed for abnormal development (e.g. heartbeat, development of eyes, etc.), and coagulated eggs can be removed easily without posing the risk of spreading bacterial or fungal infection to other eggs); however, before hatch, eggs should be placed into larger test vessel (replacement of larvae after hatch would cause stress and should be reduced to a minimum level) • if the former point is implemented, time to hatch varies from species to species and should therefore be described in this guideline (e.g., advice that fish species x should be placed in the main test vessel after y hours post-fertilisation or as soon after hatch as possible) • in a semi-static test, replacement of larvae (after hatch) with a glass tube/pipette might not be adequate as this causes stress; test medium renewal should be performed by changing approximately ¾ of the test

Category	Description
	medium as described in 'Test Solutions: 17./(ii)'; for rainbow trout, replacement with a glass tube/pipette is not adequate at all • It should be possible to choose between reporting either the stage in which the exposure of the embryos started or the time after fertilisation (e.g. < 8 hours). • Loading: The number of replicates and fish used should be reconsidered. The use of three replicates is not very common among fish tests. Two replicates with ten fish each should be sufficient. However, this may require further discussion around the power of the test to detect sub-lethal effects. • Observations, 32: 'Dead embryos and larvae should be removed as soon as observed.' Should be replaced by: 'If possible, dead embryos and larvae should be removed as soon as observed.' Sometimes it is not possible to remove coagulated eggs as they stick to each other. • Observations, 32.: for embryos: absence of heart-beat might be difficult to observe without the use of a microscope (warm-water species). • Observations, 36.: Weights: For warm water species it is hardly possible to determine individual dry weights. Therefore the additional option of group dry weights should be added.
Statistics	*Is the guidance presented in the TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* Statistical guidance is given (§ 38-40), however some points might be considered additionally. • TG does not give any advice if the main goal is the determination of a NOEC and LOEC (like in the fish early life stage) or the calculation of a LC_{50} (such as in the acute fish test) • additionally mention the William's test? • no indication of the unit of comparison (individual fish or replicate test vessel is made) • no guidance on analysis of dilution water control and solvent control
Terminology	*Is all terminology current, is the TG consistent e.g. describing life-stages etc.* Life stage terminology • limited to embryos and larvae • § 20: duration uses developmental stage (before onset of the gastrula stage); allow additional usage of time after fertilisation, e.g. < 8 hours old. Other • scientific name for zebrafish has changed from *Brachydanio rerio* to *Danio rerio*
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • TG states that a fish acute toxicity test (OECD TG 203) preferably in the test species, water solubility, vapour pressure and analytical method should be available. • Test concentrations are determined by the need to achieve a NOEC (and preferably LOEC) or a LC50 for a specific endpoint. • § 23 In general 5 concentrations should be used and justification must be provided if fewer than 5 concentrations are used. 'Concentrations of

Category	Description
	the substance higher than the 96-hour LC50 or 100 mg/L, whichever is the lower, need not be tested.' No tests above the limit of solubility. • No guidance on range finding given.
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • Minimum numbers of eggs per treatment level are given (§ 21; at least 30 eggs fertilised eggs divided equally (or as equally as possible) between 3 replicates/treatment.) • Maybe a reduction to 2 replicates with 10 eggs/each is possible. • § 32: abnormal appearance: 'Abnormal animals should only be removed from the test vessels on death'. Under certain circumstances a deformity may be so severe that the animal should be removed before death and terminated to avoid suffering. This is a requirement of some national animal welfare regulatory bodies (e.g. UK Home Office). Therefore, guidance allowing for the minimisation of suffering and meeting local legislative requirements may be helpful. • Fish Embryo Test (FET – when validated) should be recommended as a range-finder
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • Preference for non-solvent delivery is clearly stated (§ 16). • § 16 When a solubilising agent is used it must have no significant effect on survival nor visible adverse effect on the early-life stages as revealed by a solvent-only control.
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • There is no evidence that a certain species is systematically more sensitive than the other commonly used species. However, it is very rare for tests on multiple species to be performed. Testing a non-standard OECD species may be required if that species has been shown to be acutely significantly more sensitive. This is allowed for by the TG (page 3/20, Selection of fish species). • Species split into recommended (rainbow trout, fathead minnow, zebrafish, Common carp and Japanese medaka) and other well documented species (see table 1B). • Tests with rainbow trout are generally longer to perform, seasonal in terms of embryo (or gamete) availability, eggs are not translucent and vary in fertilisation success (which can only be confirmed well into the test). This should be considered opposite the advantages of using warmer water species (fathead minnow, zebrafish and Japanese medaka). • Very rare for tests on multiple species to be performed. • However, potentially testing on a non-standard OECD species may be required if that species has been shown to be significantly more acutely sensitive.
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* • Page 3/20, Handling of embryos and larvae, 14. Last sentence: ' In any case, it is recommended that handling of

Category	Description
	embryos and larvae be embryo.' Typo! • The scientific name for zebrafish has changed from *Brachydanio rerio* to *Danio rerio* since the TG was written. • § 21 loading suggests that the flow rate should be such that at least 60% of the oxygen air saturation value is achieved without aeration. This may not be achievable depending on the properties of the test substance. • The test is considered a non-animal test in many countries, provided that it ends before fish reach the free feeding stage. However, the exact point at which this occurs may be ill defined in practice. • If the test is to be considered a non-animal test, using the TG 203 as a rangefinder contradicts this. • On ethical grounds, feeding should be considered (as an option?) in the test. • For the use of solvents, and other means to dissolve or disperse the test substance, the TG should refer to the Guidance Document No. 23 • As the test allows for considerable variation in its design (e.g. number of test chambers, test concentrations, starting number of fertilised eggs), a particular test set-up should be reviewed by a statistician. • Though this TG was conditionally recommended for deletion, it should also be noted that one Member Country finds this sub-chronic fish test a valuable candidate protocol for their future implementation of effluent regulation (need to carry over some aspects to other TGs, e.g., extended FET).

6.6 OECD TG 215: Fish, Juvenile Growth Test (adopted 21st January 2000)

Category	Description
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* Chronic exposure assessing impacts on growth of juvenile fish. • Weight (mass) • Growth rates • Abnormal appearance • Abnormal behaviour (qualitative) • Survival/ mortality (%) • NOECs and LOECs (if possible) should be determined for each of the responses assessed assuming the data are amenable to statistical analysis.
Strengths	• Established methodology that has been ring tested • The TG can be applied to any of the substance types as described in Chapter II assuming the physicochemical properties of the test item allow for testing.
Limitations	*What are the limits of the data/information?* • The test does not cover all life-stages (sexual development/maturation, or the reproductive phase)
Statistics	*Is the guidance presented in the TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* Statistical guidance is given but not absolutely prescribed due to potential variations in test design (§ 43).
Terminology	*Is all terminology current, is the TG consistent e.g. describing life-stages etc.* Adequate
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • The TG states that a fish acute toxicity test (OECD TG 203) preferably in the test species, water solubility, vapour pressure and analytical method as well as biodegradability data should be available. • Test concentrations are determined by the need to achieve a NOEC and LOEC. • No guidance on range finding given.
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • No minimum numbers of fish per treatment level are given • Rather, it is suggested that a power analysis is used to determine the statistical power at which a given difference in growth rate is required to be detected. Further guidance for optimising statistical power whilst minimising the number of fish used would be beneficial.

Category	Description
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • Preference for non-solvent delivery is clearly stated • Recommendation that if a solvent is used that the actual amount should be minimised as far as practically possible is given.
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • There is no evidence that a certain species is systematically more sensitive than the other commonly used species. However, size increase for trout is typically greater than for Japanese medaka and zebrafish; therefore, trout may allow greater sensitivity in results. • There are three recommended species • tests with rainbow trout may be significantly longer than tests with other small fish species (although this test was ring tested in the trout). This should be considered opposite the advantages of using warmer water species (fathead minnow, zebrafish and Japanese medaka). • Very rare for tests on multiple species to be performed. • However, potentially testing on a non-standard OECD species may be required if that species has been shown to be acutely more sensitive.
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* – not widely used in regulation; however some researchers claim that the test is of value as it may be more sensitive than other available chronic fish toxicity TGs

6.7 OECD TG 229: Fish Short-Term Reproduction Assay (adopted Sept. 2009)

6.7.1 Purpose

The Fish Short Term Reproduction Assay (OECD TG 229) is designed to detect chemicals that affect reproductive success, including those that directly interact with the hypothalamic-pituitary-gonadal (HPG) axis. The test is conducted with adults of one of three species, the zebrafish, fathead minnow or Japanese medaka. Chemicals of interest are tested at three concentrations, and are administered via a flow-through system, ideally without use of carrier solvents. Test chemical concentrations are confirmed periodically using appropriate instrumentation during the test. Two replicate tanks containing five fish of each sex are used per treatment for zebrafish and Japanese medaka, while with the fathead minnow, four replicate tanks each with four females and two males are used per treatment. Chemical exposures are initiated after a 1-2 week acclimation/pre-exposure phase using groups of animals that, during that time, have proven to be successful spawners. The test is terminated after a 21-d chemical exposure.

Category	**Description**
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* The TG 229 protocol considers both "apical" endpoints reflective of health of the fish (i.e., survival, behaviour, fecundity, most histological changes in the gonad) and "mechanistic" endpoints indicative of specific alterations in the HPG axis (secondary sex characteristics, vitellogenin levels, some types of gonadal histopathology). Survival and qualitative assessments of appearance and behaviour, as well counts of number of eggs produced, are made daily during the pre-exposure and exposure phases of the test. At test conclusion the animals are anesthetized, and secondary sex characteristics (Japanese medaka, fathead minnow) are assessed using semi-quantitative measures. Gonad samples from the fish are removed and preserved for subsequent histological analysis (OECD 2009), and appropriate samples (plasma: fathead minnow; plasma or head/tail homogenate: zebrafish; liver: Japanese medaka) collected for vitellogenin analysis. Vitellogenin protein in the samples is measured using ELISA, with homologous antibodies and standards.
Strengths	• TG 229 has few limitations relative to substances that could be tested. (HPG-active chemicals may include inorganic (e.g., metals) and a wide range of organic substances, including pesticides, pharmaceuticals of varying physicochemical properties). • terminology used in TG 229 is appropriate and up-to-date with respect both to extant technical literature in the areas of ecotoxicology and testing endocrine-active chemicals
Limitations	*What are the limits of the data/information?* – The assay would not be suitable for very volatile chemicals (basically, those substances without the constraints of meaningful chemical delivery)
Statistics	*Is the guidance presented in the TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* – There is relatively detailed statistical guidance associated with TG 229 that was developed in consultation with

Category	Description
	several statistical consultants knowledgeable in the field.
Terminology	*Is all terminology current, is the TG consistent e.g. describing life stages etc.* – The terminology used in TG 229 is appropriate and up-to-date with respect both to extant technical literature in the areas of ecotoxicology and testing endocrine-active chemicals.
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • Given resource investments necessary for the test in terms both of time and materials, it is highly advisable to have prerequisite data to aid in setting appropriate test concentrations. There is the possibility that, for some chemicals, these data exist (e.g., via the searchable literature); however, for many of the substances tested, some sort of "range-finder" assay will be required. Guidance for achieving this (e.g., recommendations as to test length and chemical concentrations, number of animals, endpoints) is very minimal in TG229.
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • As part of development of the TG 229 protocol, power analyses were conducted to help ensure that the test was statistically robust in the context of the number of animals used. Improving the power for the endpoint fecundity with an extended TG 229 that includes sexual development is under discussion.
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* The guidance on solvent limitation is clear.
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • Three species used for this test: zebrafish, Japanese medaka and fathead minnow. • Some differences between species in terms of ease of testing (e.g., obtaining/counting eggs) and evaluation of certain endpoints (e.g., secondary sex characteristics) • Overall, the zebrafish, Japanese medaka and fathead minnow all are well-established small fish models that have been widely used in regulatory ecotoxicology. All three species can be easily cultured in the lab throughout their entire life-cycle, and are maintained and tested in many contract labs around the world.
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* • The TG 229 assay will detect chemicals with the potential to affect survival and reproductive success in fish, both important predictors of population status. There are many publications in the open literature using the basic protocol (or slight variations thereof) that have shown reproductive effects caused by chemicals with a variety of mechanisms of action (including those that likely do not act primarily through the

Category	Description
	HPG axis). However, because fecundity is a relatively variable endpoint, power of the test to detect reproductive effects could be substantially enhanced through increasing replication (and, hence, number of animals needed). This is, unfortunately, at odds with the concept of a rapid, relatively cost-effective screening assay. • Validation studies with known endocrine-active chemicals in support of development of TG 229 have shown that the mechanistic endpoints collected in conjunction with the test should effectively identify several key HPG pathways of concern: estrogen receptor agonists (increased vitellogenin in male fish), androgen receptor agonists (induction of male secondary sex characteristics in females) and inhibitors of steroid synthesis (like aromatase inhibitors) or estrogen receptor antagonists (depression in vitellogenin concentrations in females). However, there are shortcomings with TG 229 relative to identification of all HPG pathways of current regulatory concern. First, due to ambiguity relative to secondary sex characteristics, this is not a robust endpoint in zebrafish and so could affect identification of androgen receptor agonists. • The TG 229 includes guidance as to data interpretation. However, because the assay has not yet been used to support regulatory decision-making for a wide range of chemicals which could produce an unanticipated suite of responses, there undoubtedly will be an evolution in terms of how test data are interpreted. One set of responses that almost certainly will prove challenging relative to interpretation in the context of further testing will be for chemicals that affect reproductive success (fecundity) but do so without concurrently changing the two endpoints known to be directly influenced through the HPG axis—vitellogenin concentrations and secondary sex characteristics. Chemicals such as this could be affecting reproduction through HPG mechanisms not captured by discrete endpoints in the TG 229 (e.g., androgen receptor antagonists), or could be acting through non-HPG mechanisms. Gonad histology may help in differentiating the two different scenarios, but this endpoint also can be somewhat non-specific relative to reflected mechanism of action. • As the test is conducted more frequently with diverse chemicals, it seems quite likely that the section on data interpretation will need to be updated / revised. Problematic to all three species is lack of a mechanistic response that can be tied to androgen receptor antagonists, which appear to be a reactively important group of chemicals from an environmental perspective. However, these chemicals have been shown to depress fecundity in the TG 229 protocol, which is a strength the test has relative to TG 230. • The assays developed/validated through OECD for screening endocrine-active chemicals (TG 229, TG 230, TG 234, AFSS) all share the characteristic of being relatively novel in terms of design and endpoints compared to previous fish tests that have been used for regulatory ecotoxicology. This has been necessitated by the desire to identify chemicals that operate via specific a mechanism(s) of action rather than a more generic (apical) determination of toxicity. Nonetheless, this has,

Category	Description
	and will for the near future present challenges for the contract labs that normally conduct the bulk of regulatory toxicity testing. The full ramifications of this in terms of implementation of these newer fish tests are uncertain, but inevitable. In this regard challenging aspects of TG 229 include tissue dissection (including plasma isolation), ELISA measurements of vitellogenin and conducting/interpreting gonadal histopathology.

6.8 OECD TG 230: 21-day Fish Screening Assay (adopted Sept. 2009)

6.8.1 Purpose

TG 230 is similar in many regards to TG 229 (Fish Short Term Reproduction Assay), except reproductive (fecundity) data and gonadal histopathology information are not collected. The protocol focuses specifically on detecting a subset of chemicals that interact with the hypothalamic-pituitary-gonadal (HPG) axis: estrogen and androgen receptor agonists and aromatase inhibitors. The test is conducted with adults of one of three species, the zebrafish, fathead minnow or Japanese medaka. Chemicals of interest are tested at three concentrations, and are administered via a flow-through system, ideally without use of carrier solvents. Test chemical concentrations are confirmed periodically using appropriate instrumentation during the test. Two replicate tanks containing five fish of each sex are used per treatment for zebrafish and Japanese medaka, while with the fathead minnow, four replicate tanks each with four females and two males are used per treatment. Chemical exposures are initiated after a 1 week acclimation/pre-exposure phase. The test is terminated after a 21-d chemical exposure.

Category	**Description**
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* The TG 230 protocol considers survival and "mechanistic" endpoints indicative of specific alterations in the HPG axis (secondary sex characteristics, vitellogenin levels). Survival is assessed daily during the pre-exposure and exposure phases of the test. At test conclusion the animals are anesthetized, and secondary sex characteristics (Japanese medaka, fathead minnow) are assessed using semi-quantitative measures, and appropriate samples (plasma: fathead minnow; plasma or head/tail homogenate: zebrafish; liver, Japanese medaka) collected for vitellogenin analysis. Vitellogenin protein in the samples is measured using ELISA, with homologous antibodies and standards.
Strengths	• The terminology used in TG 230 is appropriate and up-to-date with respect both to extant technical literature in the areas of ecotoxicology and testing endocrine-active chemicals. • TG 230 has few limitations relative to substances that could be tested. (HPG-active chemicals may include inorganic (e.g., metals) and a wide range of organic substances, including pesticides, pharmaceuticals of varying physicochemical properties).
Limitations	*What are the limits of the data/information?* • The assay would not be suitable for very volatile chemicals (basically, those substances without the constraints of meaningful chemical delivery)
Statistics	*Is the guidance presented in the TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • There is relatively detailed statistical guidance associated with TG 230 that was developed in consultation with several statistical consultants knowledgeable in the field.
Terminology	*Is all terminology current, is the TG consistent e.g. describing life-stages etc.* • Terminology current
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.*

Category	Description
	• Given resource investments necessary for the test in terms both of time and materials, it is highly advisable to have prerequisite data to aid in setting appropriate test concentrations. There is the possibility that, for some chemicals, these data exist (e.g., via the searchable literature); however, for many of the substances tested, some sort of "range-finder" assay will be required. Guidance for achieving this (e.g., recommendations as to test length and chemical concentrations, number of animals, endpoints) is very minimal in TG 230.
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • As part of development of the TG 230 protocol, power analyses were conducted to help ensure that the test was statistically robust in the context of the number of animals used.
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* The guidance on solvent limitation is clear.
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • Three species used for this test: zebrafish, Japanese medaka and fathead minnow. • Some differences between species in terms of ease of testing (e.g., obtaining/counting eggs) and evaluation of certain endpoints (e.g., secondary sex characteristics) • Overall, the zebrafish, Japanese medaka and fathead minnow all are well-established small fish models that have been widely used in regulatory ecotoxicology. All three species can be easily cultured in the lab throughout their entire life-cycle, and are maintained and tested in many contract labs around the world.
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* • Validation studies with known endocrine-active chemicals in support of development of TG 230 have shown that the mechanistic endpoints collected in conjunction with the tests should effectively identify several key HPG pathways of concern: estrogen receptor agonists (increased vitellogenin in male fish), androgen receptor agonists (induction of male secondary sex characteristics in females) and aromatase inhibitors or estrogen receptor antagonists (depression in vitellogenin concentrations in females). However, there are shortcomings with TG 230 relative to identification of all HPG pathways of current regulatory concern. First, due to ambiguity relative to secondary sex characteristics, this is not a robust endpoint in zebrafish and so could affect identification of androgen receptor agonists. Problematic to all three species is lack of a mechanistic response that can be tied to androgen receptor antagonists, which appear to be reactively important group of chemicals from an environmental perspective. • The assays developed/validated through OECD for screening endocrine-active chemicals (TG 229, TG 230, TG 234, AFSS) all share the characteristic of being relatively novel in terms of design and

Category	Description
	endpoints compared to previous fish tests that have been used for regulatory ecotoxicology. This has been necessitated by the desire to identify chemicals that operate via specific a mechanism(s) of action rather than a more generic (apical) determination of toxicity. Nonetheless, this has, and will for the near future present challenges for the contract labs that normally conduct the bulk of regulatory toxicity testing. The full ramifications of this in terms of implementation of these newer fish tests are uncertain, but inevitable. In this regard challenging aspects of TG 230 include tissue dissection (including plasma isolation) and ELISA measurements of vitellogenin. Data Interpretation • The TG 230 includes guidance as to data interpretation. However, because the assay has not yet been used to support regulatory decision-making for a wide range of chemicals which could produce an unanticipated suite of responses, there undoubtedly will be an evolution in terms of how test data are interpreted. For example, many known endocrine-active chemicals exert effects through multiple HPG pathways. It is uncertain how the effects of these types of chemicals will be manifested in TG 230. As the test is conducted more frequently with diverse chemicals, it seems quite likely that the section on data interpretation will need to be updated / revised.

6.9 OECD TG 234: Fish Sexual Development Test (FSDT) (adopted 28 July 2011)

Category	Description
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* Chronic exposure of fish early life-stages until sexual differentiation (embryos, larvae and juveniles) assessing impacts on mortality, hatch, growth, development, sex ratio and sexual development. • Hatch success (%) • Time to hatch (time; days) • Abnormal appearance; deformity (number) • Abnormal behaviour (qualitative) • Survival/ mortality at embryo, larval, juvenile stages and overall (%) • Weight (mass) • Length; standard, fork or total • Vitellogenin (quantitative) • Sex ratio • Gonadal histopathology (optional) • Genetic sex vs. phenotypic sex in some species Effects on sex ratio, vitellogenin and certain histopathological findings[7] can be considered indicative of endocrine mediated effects. Additionally the androgen responsive protein spiggin is measured in the three-spined stickleback, and the genetic sex is determined whenever possible (e.g. in Japanese medaka and three spined stickleback) • NOECs and LOECs (if possible) should be determined for each of the responses assessed assuming the data are amenable to statistical analysis.
Strengths	• test can be applied to any of the substance types as described in Chapter 2 assuming the physicochemical properties of the chemical allow for long term testing. Generally, substances tested will be putative endocrine active substances • Established methodology that is known to provide results similar or identical to those from longer, more complex partial or full lifecycle tests. • test combines biomarkers of potential endocrine disruption and population relevant endpoints such as sex ratio. • the established methodology assesses impacts of the chemical on specific life-stages that are known to be sensitive to certain endocrine disrupting modes of action (e.g. inhibition of steroidogenesis). • Sensitive to endocrine disrupting chemicals (estrogenic, androgenic and aromatase inhibiting chemicals)
Limitations	*What are the limits of the data/information?* • The test does not cover the reproductive phase, which for certain modes

[7] Note: Specific findings including presence of testicular oocytes, Leydig cell hyperplasia, decreased yolk formation, increased spermatogonia and perifollicular hyperplasia may be considered endocrine specific – see OECD (2009). Draft OECD Guidance Document for the Diagnosis of Endocrine-Related Histopathology of Fish Gonads. Paris, Organisation for Economic Co-operation and Development

Category	Description
	of action is known to be a more sensitive stage than sexual maturation • The egg stage is resistant to the absorption of some toxicants, so the test is not as sensitive as full life cycle tests to some substances that have the ability to interfere at lower concentrations with embryonic development than with larvae or juvenile development.
Statistics	*Is the guidance presented in the TG current and in line with OECD Guidance Document No54 (2006) on Current Approaches to Statistical Analysis of Ecotoxicity Data? (Consideration of the benefits/practicalities of square root allocation.)* The test is designed for an analysis of the variance to determine a LOEC/NOEC, rather than an analysis of the regression to determine a specific effect concentration (EC_x). Guidance is clear for the proportion of sex and vitellogenin.
Terminology	*Is all terminology current, is the TG consistent e.g. describing life-stages etc.* Life stage terminology • Limited to embryos, larvae and juveniles • Duration uses developmental stage (before cleavage of the blastodisc commences, or as close as possible after this stage and no later than 12 hours post fertilization). •
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • Normally a fish acute toxicity test (OECD TG 203) preferably in the test species, water solubility, vapour pressure and analytical method should be available. Other data, if available, will be informative for the test design. Particularly the results from fish early life-stage test (OECD TG 210) and fish endocrine screening tests (OECD TG 229 or 230) • Test concentrations are determined by the need to achieve a NOEC (and preferably LOEC).
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • The number of eggs per treatment (n=120), divided between 4 replicates is the results of considerations of statistical power of the test to detect a biologically significant change in sex ratio and change in vitellogenin levels, and considerations of mortality during larval and juvenile life-stages. This should be considered as an optimal number, provided the validity criterion on mortality at the various life stages is met.).
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • Preference for non-solvent delivery is clearly stated • Validity criterion on solvent effect unclear as to what is considered adverse: '... nor produce any other adverse effects on the early-life stage as revealed by a solvent only control.' • No mention or reference to alternative preparation techniques. • There is a recommendation that if a solvent is used that the actual

Category	Description
	amount should be minimised as far as practically possible (preferably not greater than 0.1ml/L) and identical in all test chambers, except the dilution water control.
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • There is no evidence that a certain species is systematically more sensitive than the other commonly used species. • Species recommended will be limited to those for which solid validation data have been generated: zebrafish (*Danio rerio*), Japanese medaka (*Oryzias latipes*) and three-spined stickleback (*Gasterosteus aculeatus*). Fathead minnow (*Pimephales promelas*) needs to be validated with an androgen agonist. • Other fish species may be used provided their biological development, and in particular the period of sexual development, is sufficiently well known. Proficiency chemicals might be recommended. • Japanese medaka and the three-spined stickleback present the additional advantage that their genetic sex can be determined in addition to their phenotypic sex.
General	*Are there any points of clarity/interpretation required following experience of the TG in regulatory application?* • Limited experience on the regulatory application of the test makes it difficult to judge on clarity of the test and interpretation. One potential issue concerns interpretation of data and derivation of NOEC/LOEC on population relevant endpoints vs. biomarker endpoints? (e.g. can a NOEC/LOEC be derived based on vitellogenin levels if most sensitive endpoint despite not being a population relevant endpoint?)

6.10 Androgenised Female Stickleback Screen (AFSS) (published 18 August 2011, No. 148 in the Series on Testing and Assessment, ENV/JM/MONO(2011)29)

6.10.1 Purpose

The androgenised female stickleback screen (AFSS) is designed to identify chemicals that interact with the androgen receptor, especially as antagonists. Identification of test chemicals as androgen receptor antagonists is based on their ability to block the biological activity (induction of spiggin) of a model androgen receptor agonist dihydrotestosterone (DHT) in female adult sticklebacks. The test includes seven treatment groups: (1) a water-only control, (2) solvent control, (3) negative control (test chemical at a "high" concentration), (4) positive control (5 µg DHT/L), (5) high test chemical concentration plus DHT, (6) medium test chemical concentration plus DHT, and (7) low test chemical concentration plus DHT. As opposed to TGs 229 and 230, solvent use cannot be totally avoided in the test since DHT requires a carrier. Test chemical concentrations are confirmed periodically using appropriate instrumentation during the test. Two replicate tanks containing five females are used per treatment. Chemical exposures are initiated after a 1 week acclimation period, and the test is terminated after a 21-d chemical exposure.

Category	**Description**
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* Survival and qualitative assessments of appearance and behaviour are made daily during the pre-exposure and exposure phases of the AFSS test. At test conclusion the animals are anesthetized, and the kidney is removed for spiggin analysis. Spiggin, a protein normally produced only in male sticklebacks, is used as cementing material for nest construction. However, exposure of females to exogenous androgen receptor agonists (such as DHT) stimulates spiggin production. Inhibition of this stimulation suggests that a test chemical may be an androgen receptor antagonist. Spiggin is measured using an ELISA
Strengths	• The AFSS assay, like TGs 229 and 230, was developed specifically to detect HPG-active chemicals. These chemicals could include both inorganic (e.g., metals) and a wide range of organic substances, including pesticides, pharmaceuticals and high-production volume chemicals, of varying physicochemical properties. Basically, within the constraints of meaningful chemical delivery (e.g., the assay would not be suitable for very volatile chemicals), the AFSS has few limitations relative to substances that could be tested. The model androgen DHT is used as part of the AFSS design • The AFSS will detect chemicals that inhibit DHT-induced spiggin production in female sticklebacks. Several published studies have shown that one important class of endocrine-active chemicals that will do this is androgen receptor antagonists. In this regard the AFSS fills a niche not covered by TGs 229 and 230, neither of which have mechanistic endpoints that serve to specifically identify androgen receptor antagonists
Limitations	*What are the limits of the data/information?* • The assays developed/validated through OECD for screening endocrine-active chemicals (TG 229, TG 230, AFSS) all share the characteristic of being relatively novel in terms of design and endpoints compared to previous fish tests that have been used for regulatory ecotoxicology. This has been necessitated by the desire to identify

Category	Description
	chemicals that operate via specific a mechanism(s) of action rather than a more generic (apical) determination of toxicity. Nonetheless, this does, and will for the near future, present challenges for the contract labs that normally conduct the bulk of regulatory toxicity testing. The full ramifications of this in terms of implementation of these newer fish tests are uncertain, but inevitable. In this regard challenging aspects of the AFSS include the dual chemical exposures, kidney dissection and spiggin ELISA measurements.
Statistics	*Is the guidance presented in the GD current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • There is relatively detailed statistical guidance associated with AFSS protocol that was developed in consultation with several statistical consultants knowledgeable in the field.
Terminology	*Is all terminology current, is the GD consistent e.g. describing life stages etc.* • The terminology used in the AFSS GD is appropriate and up-to-date with respect both to extant technical literature in the areas of ecotoxicology and testing endocrine-active chemicals.
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the GD consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • Given resource investments necessary for the test in terms both of time and materials, it is highly advisable to have prerequisite data to aid in setting appropriate test concentrations. There is the possibility that, for some chemicals, these data exist (e.g., via the searchable literature); however, for many of the substances tested, some sort of "range-finder" assay will be required. Guidance for achieving this (e.g., recommendations as to test length and chemical concentrations, number of animals, endpoints) is very minimal in the AFSS GD.
Animal minimisation	*Does the GD sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • As part of development of the AFSS protocol, power analyses were conducted to help ensure that the test was statistically robust in the context of the number of animals used
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • Use of solvent is unavoidable as it is required as a carrier for DHT
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • The three-spined stickleback is used for the AFSS test. This species, while not commonly used in the past for ecotoxicology, is gaining favour as an experimental model, in part because its genome has been sequenced. In the past much of the experimental work done with the stickleback had been with field-collected (wild) fish, but there are now viable approaches for culturing this species in the lab, thereby enhancing its value as a toxicological model.
General	*Are there any points of clarity/interpretation required following experience of the GD in regulatory application?* • The AFSS protocol includes guidance as to data interpretation.

Category	Description
	However, because the assay has not yet been used to support regulatory decision-making for a wide range of chemicals which could produce unanticipated responses, there undoubtedly will be an evolution in terms of how test data are interpreted. For example, it is possible that endocrine-active chemicals other than androgen receptor antagonists could inhibit DHT-induced spiggin production. Recent studies analogous to the AFSS in the fathead minnow have shown that, in addition to androgen receptor antagonists, estrogen receptor agonists can modulate (inhibit) the masculinizing effects of androgens in female fish. • As the test is conducted more frequently with diverse chemicals, it seems quite likely that the section on data interpretation will need to be updated or revised.

6.11 OECD TG 305: Bioconcentration: Flow-through Fish Test (adopted 14th June 1996)

This Test Guideline is currently being revised to include the possibility of reducing the cost and number of laboratory animals used, when this can be done without compromising the BCF determination. The revision also includes a possibility to estimate a bioaccumulation factor (BAF) from dietary exposure of the fish, when such a test design is warranted, because the high hydrophobicity of the substance implies difficulties in exposing the fish *via* water. A dietary ring test was performed.

Category	**Description**
Deliverables	The purpose of the test is to investigate the uptake and depuration of a test chemical in fish. Fish are exposed to the chemical in solution. Main deliverable is the bioconcentration factor (BCF), at steady state and/or kinetic based on whole fish To calculate the BCF, the following information is measured at several time points: • Test chemical concentration in water (mean measured) • Test chemical concentration in fish tissue • Lipid content of fish • Sampled fish weight (mass) Results can also be determined for specific fish tissues (edible (fillet) and non-edible (viscera) fractions)
Strengths	• Established method that is known to provide robust results for a wide range of organic substances, accepted for regulatory use worldwide. Fulfils needs of risk assessment (secondary poisoning), PBT assessment and classification and labelling. • suitable for a range of stable organic chemicals with a log Kow in the range 1.5 - 6. "Super lipophilics" with a log Kow >6 may be tested in some cases.
Limitations	*What are the limits of the data/information?* • Not well suited for unstable substances, poorly water soluble/highly lipophilic substances, surfactants, strongly adsorbing substances, complex mixtures • No clear guidance on how to euthanise fish • Bioconcentration tests are not well suited for complex substances (eg UVCBs), where uptake may occur for some components but not for others. In such cases the basis of the analytical technique is very important to identify what is being accumulated. • The measurement of accurate aqueous test substance concentrations is very important. The presence of undissolved test substance affects bioavailability, and if the method for analysing concentrations in water cannot distinguish between the truly dissolved fraction and emulsions, then the (steady state) BCF may be underestimated. Furthermore, disparities between the sensitivity and specificity of the analytical techniques for water and fish concentrations can have implications for results. • It is assumed that in most cases first order kinetics will be followed. This may be the case for the majority of substances, but for those that

Category	Description
	differ little guidance on how to treat their data is given (important for when an apparent steady state has not been reached). • The definition of what steady state is in the study is not given until the definitions part in Annex 1 – it should perhaps appear earlier. • The guideline makes several references to the use of complex models to interpret kinetic data that does not adhere to (roughly) first order, but does not go on to describe how any such models might be used (one reference is given to the Spacie and Hamelink 1982 paper) . • Lipid normalisation: the TG suggests that results should be based on total lipids for substances with log Kow >3. Details of how to do this, or how to interpret such a result, are not given. • differences in metabolic pathways for different species of fish may impact on interspecies BCFs • result gives limited information on possibility of biomagnification and trophic transfer
Statistics	*Is the guidance presented in the draft TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • No statistical guidance given but importance recognised for comparison of test group data with control data (and identification of "outliers"); for statistical power in results (effect of varying number of sampling points/number of fish sampled at each sampling point); and effect that pooling data has on statistical power • The test differs from the other TGs with respect to statistics in its design.
Terminology	*Is all terminology current, is the draft TG consistent e.g. describing lifestages etc.* • Some of the definitions in annex 1 need updating

Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the draft TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.*
	• The TG lists seven pieces of prerequisite information along with the OECD guideline that can be used for them
	  • solubility in water
	  • octanol-water partition coefficient
	  • hydrolysis
	  • phototransformation in water
	  • surface tension
	  • vapour pressure
	  • ready biodegradability
	• Although not listed with the above, the TG goes on to say that a suitable analytical technique for the test substance should be available, capable of measuring at least a tenfold decrease in concentration from the concentration tested.
	  • Adequate information given on concentration setting:
	  o take into account toxicity data (higher test concentration to be 1% of acute asymptotic LC50 and tenfold higher than the limit of detection
	  o second (and subsequent) concentration(s) to differ by a factor of

Category	Description
	ten, but maybe less depending on analytical LoD and toxicity o No concentration to be above water solubility limit
Animal minimisation	*Does the TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* **Reduction** *Number of test concentrations:* The TG requires that *at least* three groups of fish are included in a study (a control, low concentration and high concentration test groups). Avoiding the use of further test concentration groups will have the greatest effect on the number of animals used in the test. Experience gained with the test shows that in the majority of cases one concentration may be sufficient (since approximate first order kinetics are usually followed) **Refinement** *Study length/number of sampling points:* Prediction of the length of the uptake and duration phases is recommended before a test to plan the sampling schedule and identify how many animals may be required accordingly. The TG recommends that at least five sampling points in the uptake phase and four in the depuration phase are employed. For longer studies (up to a maximum of 60 days uptake, no limit given for depuration duration) more sampling points are likely to be needed. It is left up to the experimenter to plan their sampling schedule (although analytical costs as well as animal minimisation may presumably be a strong driver for keeping sampling points to a minimum). *Number of fish sampled:* the TG requires at least four fish to be sampled from each group at each sampling point. Increasing this number can be considered if greater statistical power is needed. In both these cases the TG does not go on to say that efforts should be made to keep the number of animals used to a minimum whilst ensuring that the data generated is of sufficient quality for its purpose. *Lipid analysis:* the TG states that it is preferable that lipid analysis is carried out on the fish sampled for test concentrations (this benefits both the results' accuracy and reduction in fish numbers). However the number of fish saved in this way is likely to be low, as typically only about six fish are sampled for lipid in a study.
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • Preference for non-solvent/non-dispersant delivery clearly stated • If used, a solvent control is also necessary
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • Species split into those recommended (freshwater, temperate and tropical) and other species (marine, estuarine) which have been used (see Annex 3). • In practice, rainbow trout, zebrafish and common carp have been widely used • TG does not recommend tests on multiple species • Testing on a "non-standard" OECD species is not ruled out in the TG
General	*Are there any points of clarity/interpretation required following experience of the draft TG in regulatory application?* • the title is misleading in that semi-static conditions are allowed • it is stated that various pieces of information on the test substance

Category	Description
	(according to other OECD TGs) *should* be available before carrying out a test; however some of these in practice are really essential before conducting a test (e.g. water solubility, Kow, surface tension) while others are not (although still important nonetheless, e.g. phototransformation in water, biodegradability) • Details relating to the use and interpretation of results for radio-labelled test substances are not so clear. The guideline states that metabolites "may be characterised if deemed necessary."… "BCFs based on total radio-labelled residues can serve as one criterion for determining if degradates identification and quantification is necessary", although a criterion is given ("may be advisable" if BCF ≥1000 to quantify degradates representative of ≥10% of total residues at steady state). Overall the text is not explicit when such identification should be done, possibly because it is reliant on why the study is being conducted. • The guideline states that a depuration phase is always necessary unless uptake has been insignificant (BCF <10). This may however be misleading for the more hydrophobic test substances, where the rate of uptake from the test medium may be very slow indeed such that at the end of a standard 28 day exposure period the apparent steady state BCF may be very low, but concentrations in the fish would continue to increase significantly should the exposure period be extended. • The TG states that three consecutive sampling points must give a test substance concentration in fish within 20% for (apparent) steady state to have been reached at the end of the uptake phase. If so, then a steady state BCF can be calculated, or a BCF at a percentage of steady state taking into account the shape of the uptake curve (TG gives 80% or 95%). The TG also says that a kinetic BCF can be calculated from the uptake and depuration rate constants k_1 and k_2, but guidance on when one BCF might be preferred to the other, or possible reasons for differences between the two, is not given. • The range over which the pH of the test medium might vary during a study is given as a recommendation (± 0.5), but not included as a test validity criterion, which may add some confusion. Similarly, fish lipid content should not vary greater than ±25% during a study (but not given as a test validity criterion). • Numbers of test fish: it is not so clear that more fish may be required for longer exposure periods in the case that steady state has not been reached after 28 days. • It is a TG requirement to record any adverse effects/abnormalities observed in fish. However, the TG lacks examples of what these might be. • Some further background information relating to k1 and k2 might be useful in annex 1. • Annex 6: some of the methods here are now outdated, and may need to be modernised • The TG does not differentiate between juvenile and adult test animals. This may be important because it is known that fish size influences rate of uptake via passive diffusion at the gill; the larger the fish, the lower the rate of uptake. In the case of growing fish this may mean that a

Category	Description
	"true" steady state is not reached in the uptake phase. There is also the issue of "growth dilution" correction (see below) • flow-through conditions are not always easy to maintain, and fluctuations can cause serious problems for the TG's validity criteria (dissolved oxygen concentrations, maintenance of aqueous test substance concentrations and minimisation of DOC content). • The TG strongly discourages the use of solvents and dispersants. It is now widely accepted that solvents, up to a specific concentration limit, can be used in aquatic testing so long as the test substance is present below its solubility limit in water. The use of dispersants has been widely criticised, because of their influence on bioavailability. • The method used to measure lipid content can have a marked effect on the result. The TG recommends one method, but could give more guidance in this area. • The current guideline compares well with other related guidelines (e.g. ASTM E1022-94; ASTM 2003 and OPPTS 850.1730; US EPA 1996), but there are a number of differences (i.e. method of test water supply (other methods allow static, semi-static or flow through); other methods do not always require a depuration phase; mathematical method for calculating BCF; sampling frequency, including number of measurements in water and number of samples of fish; guidance for measuring the lipid content of the fish (for the purposes of lipid normalisation); minimum duration of the uptake phase) • Annex 6 gives details of how to determine k_1 and k_2 for calculation of a kinetic BCF. k_1 and k_2 can be calculated sequentially or simultaneously. The TG states a preference for the sequential approach, but the "better" method is open to debate and often one method will work well for one study and the other will work better for another study (by comparison of estimated curves with the measured data). • For substances with a log K_{OW} >3 it is recommended that results are presented on a total lipid content basis. This is a way of removing one area of bias in studies, and allows comparisons to be made between studies. However, how to do this is not covered. • Uptake and depuration rate constants between the tested concentrations should not differ by more than ±20% otherwise first order kinetics may not have been followed. Guidance is not given on how to account for these differences, or how to interpret the results. • There is no indication in the TG of what constitutes adequate statistical power. In the context of paragraph 28, power probably refers to the likelihood of detecting non-steady state. • No guidance on how to handle "no-detects" is given. • "Growth dilution" in studies which use juvenile fish is not referred to in the TG. Growth dilution has been recognised as a contributor to the overall depuration rate k_2 although it is not actually a removal mechanism; concentrations in fish appear lower because the fish is increasing in size. In some cases the effect can have a marked effect on measured concentrations, especially during the depuration phase. Kinetic BCFs have started to be derived routinely using growth dilution corrected depuration rate constants to account for this effect. (NB: no

Category	Description
	equivalent correction is carried out for steady state BCFs currently, although growth may have an effect on steady state, see section 9). • The conduct of a range-finding minimised chronic test, with very few fish over 28 days, might be useful in concentration setting. • In 2008 a proposal to revise the guideline was submitted to the OECD by the Netherlands, Germany and the UK. This proposal to refine the existing method to use fewer animals, add a "minimised design" test (same duration but fewer sampling points), and add a dietary method for testing very poorly soluble/highly lipophilic substances for which the existing method is unsuitable was accepted.

6.12 Zebrafish Embryo Toxicity Test (ZFET; protocol as of 13th November 2009)

Note: This compilation is based on version 2.9 of the SOP for the Zebrafish Embryo Toxicity Test (ZFET), which was used for phase 1b of the OECD Validation study and the draft test guideline provided by the lead country Germany.

Category	Description
Deliverables	*What data/information does the TG deliver (acute/chronic, endpoints etc).* Acute exposure assessing impacts on mortality. • acute toxicity of substance to fish embryos after 24, 48, 72 and 96 h (determination of LC_{50} value) • lethal effects defined by four apical observations: (1) coagulation of the embryo, (2) non-detachment of tail, (3) non-formation of somites and (4) non-detection of heart beat • designed as an alternative test method to the acute toxicity tests with juvenile and adult fish, i.e., the OECD TG 203, thus providing a reduction/replacement in fish usage
Prerequisites	Information on the test substance: • physicochemical data of test substance including water solubility, stability and biodegradability (OECD TG 301) • reliable analytical method of chemical analysis of test concentrations
Strengths	• internal and external negative controls • dilution water composition following OECD TG 203 • permanent use of a positive control (at present: 3,4-dichloroaniline) • alternative test method to the acute toxicity tests with juvenile and adult fish (OECD TG 203) • clear definition of replication • precise definition of maintenance conditions for brood stock (parental fish) and test organisms (embryos) • guidance about the use of fish previously treated against disease: extension of non-treatment period to 2 months • recommendation for semi-static renewal procedure • at least within validation study, obligatory use of pre-set reporting templates • statistical analysis following OECD GD 54 (see Statistics)
Spectrum of test substances	• generally, no restriction • TG can be applied to any of the substance types as described in Chapter II assuming the physicochemical properties of the test item allow for long-term testing
Limitations	*What are the limits of the data/information?* • substances may cause delayed hatch beyond 96 hours, which will preclude the exposure of eleutheroembryos; in cases when chemical exposure after hatch seems indispensable, other tests, e.g. OECD TG 203, might be performed; known examples of substances requiring prolonged exposure to the eleutheroembryos stage are, e.g., quaternary ammonium salts • test species recommended (at present: zebrafish, *Danio rerio*) only cover freshwater species; indication that modified SOP works for other

Category	Description
	fish species such as fathead minnow (*Pimephales promelas*) and Japanese medaka (*Oryzias latipes*) • some substances may be much less toxic to fish embryos/larvae than to juveniles/adults; such outliers are subject to investigation during the on-going validation of the ZFET; discussion required on suitability of FET as an alternative to TG 203, let alone as a replacement (risk that ZFET instead of TG 203 would miss some substances with high acute hazard to fish) • The egg stage is resistant to the absorption of some toxicants, so the test is not as sensitive as full life cycle tests to some substances that have the ability to interfere at lower concentrations with embryonic development than with larvae or juvenile development.
Statistics	*Is the guidance presented in the draft TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • during the validation period, 20 embryos per test concentration and controls are being used; depending on outcome of statistical analysis, final TG will most likely be limited to 10 embryos per concentration • various statistical methods are being explored within the validation study following OECD guidance document 54 on statistical analysis of ecotoxicity data; final version of TG will provide suggestions as to appropriate statistical methods
Terminology	*Is all terminology current, is the draft TG consistent e.g. describing life stages etc.* • precise description of life stage to be used in the test (test initiation at latest 1 h after fertilization)
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the draft TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • limit test concentration in final TG will be 100 mg/L (full test, if any mortality in the limit test) • normally 5 concentrations; justification required if fewer than five concentrations are used • in the validation study, precise guidance about concentrations of test substances preparation and controls is given in separate trial plans (for each phase of the validation study); guidance on range finding strategies might be useful for the final TG • pre-saturation of well plates with test concentrations
Quality assurance	*Which criteria are required as validity criteria?* • requirements for reproductive performance (fecundity, standard fertility rate) of parental fish (brood stock): fertility rate of the parent generation should be ≥ 70% • the water temperature should be maintained at 26 ± 1 °C in test chambers at any time during the test. • in order to control quality of brood stock, LC_{50} of the standard positive control 3,4-dichloroaniline should be routinely determined in embryos (LC_{50} between 1.6 and 4.4 mg/L) • alternatively: testing of a single fixed concentration of 3,4-dichloraniline at 4 mg/L: minimum mortality of 30 % after 96 h • overall survival of embryos in the negative external control and, where

Category	Description
	relevant, in the solvent control should be ≥ 90% until the end of exposure. • measured concentrations should be within ± 20 % deviation from nominal concentrations; if > 20 % deviation, results should be given with reference to measured concentrations • dissolved oxygen ≥ 80 % for maintenance of brood stock and during the test • semi-static renewal procedure
Animal minimisation	*Does the draft TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • alternative test method to the acute toxicity tests with juvenile and adult fish, i.e., the OECD TG 203 • if not regarded as replacement, it can at least contribute to reduction / refinement, since it might reduce the overall number of fish used for acute toxicity testing and testing is restricted to the least developed life stages possible (discussion on-going)
Non-solvent delivery / Solvent use	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?* • reference to OECD Guidance Document No. 23[8] • in the final TG, a clear recommendation for the preference for non-solvent delivery will be given • reference to solvents listed in OECD TG 215; in addition, DMSO accepted • at present, no clear recommendation that if a solvent is used that the actual amount should be minimised as far as practically possible • recommendation that solvent concentration should be identical in all test concentrations • definition of maximum solvent concentration of 1000 µl/L
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • at present, SOP specifically addresses the needs of zebrafish (*Danio rerio*) development • there is evidence that species-specific modification of the SOP allows use of other common OECD test species such as fathead minnow (*Pimephales promelas*) and Japanese medaka (*Oryzias latipes*)[9]
General	*Are there any points of clarity/interpretation required following experience of the draft TG in regulatory application?* • at an international level, clarification needed whether ZFET should be classified as replacement method or as reduction / refinement method • recommendation that consideration of the substances not suited to this test (i.e., not crossing chorion) is given (annex?)

[8] OECD (2000) OECD Series on Testing and Assessment no. 23 Guidance document on aquatic toxicity testing of difficult substances and mixtures.

[9] Braunbeck, T., Böttcher, M., Hollert, H., Kosmehl, T., Lammer, E. Leist, E., Rudolf, R., Seitz, N. (2005) Towards an alternative for the acute fish LC_{50} test in chemical assessment: the fish embryo toxicity test goes multi-species B an update. ALTEX 22: 87-102.

Category	Description
	• possibly use the ZFET as a rangefinder for TG 203 and TGs to generate more data (about 150 substances completed); efforts should be made to identify relevant physicochemical properties to give an applicability range • applicability of the ZFET as a replacement or a refinement under given national animal welfare regulations needs to be clarified

6.13 Fish Full Life-Cycle Test Guideline (FLCT; Japan)

Note: The Test Guideline has not yet been adopted. This review is based on the draft of the potential Test Guideline**.** The FLCT has been proposed as a new Test Guideline along with the Japanese medaka Multigeneration Test. As part of the validation effort of the MMT, the added value of the MMT in comparison to the FLCT is being evaluated. Once this validation effort has been completed, the addition of this FLCT and/or Japanese medaka Multigeneration Test to the suite of OECD Test Guidelines can be considered.

6.13.1 Purpose

This draft Test Guideline, based on the Japanese medaka full life-cycle test guideline developed by Japan, describes a fish toxicity test that can be used to evaluate the potential chronic effects of chemicals on fish populations. The method gives primary emphasis to potential population relevant effects (namely, adverse impacts on survival, development, growth and reproduction) for the calculation of the No-Observed Effect Concentration (NOEC). These effect observations should be augmented by secondary mechanistic biomarker responses (namely, vitellogenin, gonad somatic index [GSI], and gonad histology). The method is applicable to a variety of chemicals, including endocrine disrupters and general toxicants. The Japanese medaka (*Oryzias latipes*) is a suitable species for use in this test guideline; however, other species such as fathead minnow (*Pimephales promelas*), sheepshead minnow (*Cyprinodon variegatus*), three spined stickleback (*Gasterosteus aculeatus*) and zebrafish (*Danio rerio*) are also suitable.

Category	**Description**
Deliverables	*What data/information does the draft TG deliver (acute/chronic, endpoints etc).* The FLCT includes several endpoints at different life-stages, including embryological development, hatching (hatchability and time to hatch), post-hatch survival, growth (total length and body weight), sexual differentiation (secondary sex characteristics and gonadal histology) and hepatic vitellogenin (VTG) for both the F0 and F1 animals and reproduction (fecundity and fertility) and gonadosomatic index (GSI) for the F0 adults. Although not included in the draft guideline, a genetic sex marker is available for Japanese medaka and this could be utilized for another endpoint for assessing potential disruptions in sexual development. The genetic sex determination should be mandatory for the species where possible because it gives valuable information about phenotypic reversal which is a population relevant endpoint (see also MMT).
Strengths	– The draft TG can be applied to any of the substance types as described in Chapter 2 assuming the physicochemical properties of the test item allow for long term testing. – The FLCT is a chronic laboratory toxicity test designed to comprehensively evaluate over two generations the adverse effect threshold for individual and population relevant endpoints. In addition to the traditional chronic effect measures of survival, growth, and reproduction endpoints, the FLCT offers the ability to address both structural and activational endocrine pathways in all life stages. – Another strength and/or weakness, depending on perspective, is that the test incorporates a NOEC/LOEC statistical design as opposed to an EC_x

Category	Description
	design. There are pros and cons which are well discussed in OECD (2006) to both general designs, but in the case of the FLCT, and generally for multiple endpoint and multiple life-stage tests, the NOEC/LOEC is more logistically practical.
Limitations	*What are the limits of the data/information?* • The FLCT in-life phase is designed to be completed in approximately 26 weeks which is slightly longer than the 24 week duration of the Japanese medaka Multigeneration Test (MMT). The method is subject to all the problems of extended term testing, such as diluter malfunctions, microbial growth, mishandling, etc., that may arise over time and compromise the integrity of a test.
Statistics	*Is the guidance presented in the draft TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • There is limited statistical guidance associated with the draft FLCT and the test guideline could be greatly improved with the addition of more detailed guidance.
Terminology	*Is all terminology current, is the draft TG consistent e.g. describing life-stages etc.* • The terminology used in the draft FLCT is appropriate and up-to-date with respect both to extant technical literature in the areas of ecotoxicology and testing endocrine-active chemicals.
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the draft TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • It is expected that fish acute toxicity test (OECD TG 203) preferably in the test species, water solubility, vapour pressure and analytical method should be available. In addition, the draft FLCT guideline recommends that results from tests in the EDTA level 1 and 2, subchronic toxicity, and a range-finding test under the same conditions as the definitive test should be used to establish the appropriate test concentrations.
Animal minimisation	*Does the draft TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • It is not clear what has been done to minimize and optimize the efficient use of the animals employed in this test. If power analyses do not exist, there are a sufficient number of tests that could be utilized to confirm or establish the most appropriate design.
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?*
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • This draft FLCT guideline was prepared for Japanese medaka (*Oryzias latipes*), however, other species such as fathead minnow (*Pimephales promelas*), sheepshead minnow (*Cyprinodon variegatus*), three spined stickleback (*Gasterosteus aculeatus*) and zebrafish (*Danio rerio*) are also suitable.

Category	Description
General	*Are there any points of clarity/interpretation required following experience of the draft TG in regulatory application?* • Data interpretation guidance is lacking in this draft Test Guideline

6.14 Japanese medaka Multigeneration Test (MMT; Japan)

Note: The guideline has not yet been adopted. This review is based on the current draft of the potential test guideline. The Japanese medaka Multigeneration Test is still in the validation phase and is only proposed at this time. Once the validation program is successfully completed and peer reviewed, the addition of this method to the suite of OECD Test Guidelines can be considered.

6.14.1 Purpose

The Japanese medaka Multigeneration Test (MMT) is a proposed Test Guideline that can be used to evaluate the potential chronic effects of chemicals on fish populations. The method gives primary emphasis to potential population relevant effects (namely, adverse impacts on survival, development, growth and reproduction) for the calculation of a No-Observed Effect Concentration (NOEC). These effect observations should be augmented by secondary mechanistic biomarker responses (namely, vitellogenin, gonad somatic index [GSI], and gonad histology). The method is applicable to a variety of chemicals, including endocrine disrupters and general toxicants.

<table>
<tr><th>Category</th><th>Description</th></tr>
<tr><td>Deliverables</td><td>What data/information does the draft TG deliver (acute/chronic, endpoints etc).
The MMT provides data that can be used to simultaneously evaluate two general types of adverse outcome pathways (AOPs) ending in reproductive impairment: a) endocrine-mediated pathways involving disruption of the hypothalamus-pituitary-gonadal (HPG) endocrine axis; and, b) pathways that cause reductions in survival and growth through non-endocrine mediated toxicity. Test data provides information for evaluating adverse outcomes from either endocrine-mediated or non-endocrine mediated effects, or both. The MMT endpoint data are listed in Table 1. Some of the EDC endpoints, such as the presence of anal fin papillae in Japanese medaka males, are effect biomarkers only minimally linked to adverse reproductive outcomes; whereas other EDC endpoints such as fecundity and fertility can be directly linked to adverse outcomes through life-history translational models and, with possible additional links, to population models. Endpoints typically measured in chronic toxicity tests such as the full life-cycle test and the early life-stage (ELS) test are also included in the MMT and can be used to evaluate the hazards posed by both non-endocrine mediated toxic modes of action and endocrine-mediated toxicity pathways.

Endpoint overview of the MMT:
<table>
<tr><th>Life-stage</th><th>Endpoint</th><th>Endocrine-specific</th><th>Non-endocrine-mediated</th><th>Direct population relevance</th></tr>
<tr><td>ELS (9-14 dpf)</td><td>Hatch</td><td></td><td>●</td><td>●</td></tr>
<tr><td>ELS (2 dpf)</td><td>Survival</td><td></td><td>●</td><td>●</td></tr>
<tr><td>Su -adult (8 wpf)</td><td>Survival</td><td></td><td>●</td><td>●</td></tr>
<tr><td></td><td>Growth (weig t)</td><td></td><td>●</td><td></td></tr>
<tr><td></td><td>Gonad phenotype/Genetic sex</td><td>●</td><td></td><td>●</td></tr>
<tr><td></td><td>Vitellogenin (Vtg)</td><td>●</td><td></td><td></td></tr>
</table></td></tr>
</table>

Category	Description				
		Anal fin papillae	●		
	Adult (11-14 wpf)	Fecundity	●	●	●
		Fertility	●	●	●
	Adult (14 wpf)	Survival		●	●
		Growth (weight)		●	
		Gonad phenotype/Genetic sex	●		●
		Anal fin papillae	●		
		Histopathology			
		Gonad	●	●	
		Liver		●	
		Kidney		●	
		Other		●	
Strengths	• The draft TG can be applied to any of the substance types as described in Chapter 2 assuming the physicochemical properties of the test item allow for long term testing. • The MMT is a chronic laboratory toxicity test designed to comprehensively evaluate through multiple generations the adverse effect threshold for individual and population relevant endpoints. As such, it represents in general the ultimate in a laboratory test for evaluating chronic toxicity outcomes in fish for use in an ecological risk assessment. In addition to the traditional chronic effect measures of survival, growth, and reproduction endpoints, it offers the ability to address possible trans-generational effects of sex ratio alterations and eventual reproductive performance of offspring from exposed parents. Thus the assay allows evaluation through all life stages of both structural and activational endocrine pathways within and across generations. • The MMT in-life phase is designed to be completed in 24 weeks which is as short or shorter duration than comparable full life-cycle tests in other fish species which expose only two generations. • Another strength and/or weakness, depending on perspective, is that the test incorporates a NOEC/LOEC statistical design as opposed to an ECx design. There are pros and cons which are well discussed in OECD (2006) to both general designs, but in the case of the MMT, and generally for multiple endpoint and multiple life-stage tests, the NOEC/LOEC is the more logistically practical.				
Limitations	*What are the limits of the data/information?* • The method is still subject to all the problems of extended term testing, such as diluter malfunctions, microbial growth, mishandling, etc., which may arise over time and compromise the integrity of a test.				
Statistics	*Is the guidance presented in the draft TG current and in line with OECD 2006? (Consideration of the benefits/practicalities of square root allocation.)* • There is relatively detailed statistical guidance associated with the MMT that was developed in consultation with several statistical consultants knowledgeable in the field. In addition, the Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (OECD 2006) document is referenced as an additional guidance source..				
Terminology	*Is all terminology current, is the draft TG consistent e.g. describing life-stages etc.*				

Category	Description
	• The terminology used in the MMT is appropriate and up-to-date with respect both to extant technical literature in the areas of ecotoxicology and testing endocrine-active chemicals.
Concentration setting	*Is there sufficient guidance on how to choose test concentrations, is the draft TG consistent in respect to limit concentrations, should guidance on potential range finding strategies be elucidated.* • It is expected that a fish acute toxicity test (OECD TG 203) preferably in the test species, water solubility, vapour pressure and analytical method should be available. In addition, the MMT guideline recommends that a range-finding test under the same conditions as the definitive test be performed to establish the appropriate test concentrations.
Animal minimisation	*Does the draft TG sufficiently direct the test design to be in line with the 3Rs? Is there guidance that can be added to enhance this perspective?* • As part of development of the MMT protocol, statistician consultation and power analyses were conducted to help ensure that the test was statistically robust in the context of the number of animals used.
Non-solvent delivery	*Is the guidance on solvent limitation clear? Is there a need to develop specific guidance for individual or all aquatic TGs?*
Species effectiveness	*Is it justified to have all species as potential test organisms? When are tests on multiple species required?* • The Japanese medaka (*Oryzias latipes*) is the appropriate species for use in this draft test guideline.
General	*Are there any points of clarity/interpretation required following experience of the draft TG in regulatory application?* • Data interpretation guidance is lacking in the draft proposed Test Guideline. For this test method, data interpretation has two purposes – 1) determination of statistically supported effect levels (*e.g.,* NOEC/LOEC) and 2) interpretation relevant to endocrine disrupting activity and potential population impact. Guidance is provided for appropriate statistical analysis but the Test Guideline could be improved with discussion of endocrine relevant adverse outcome pathways (AOPs) and population modelling approaches which can be used to interpret the implications of the effects statistically resolved.

6.15 References

McKim, J.M. (1977). Evaluation of tests with early life stages of fish for predicting long-term toxicity. J. Fish. Res. Bd Can. 34: 1148-1154.

OECD (1984). Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. Fish prolonged toxicity test, 14-day study, TG 204. OECD, Paris.

OECD (1992). Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. Fish acute toxicity test, TG 203. OECD, Paris.

OECD (1992). Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. Fish early life-stage toxicity test, TG 210. OECD, Paris.

OECD (1998). Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. Fish short-term toxicity test on embryo and sac-fry stages, TG 212. OECD, Paris. 20 pp.

OECD (2000). Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. Fish juvenile growth test, TG 215. OECD, Paris.

OECD (2002) Fish two-generation test guideline. Draft proposal for a new guideline. Organisation for Economic Cooperation and Development, Paris, 18 pp.

OECD (2006) Current Approaches in the Statistical Analysis of Ecotoxicity Data: a guidance to application. OECD Series on Testing and Assessment. Guidance Document No. 54. Organisation for Economic Cooperation and Development, Paris, 146 pp.

OECD (2006) Draft Proposal for a New Guideline, Fish Embryo Toxicity (FET) Test. OECD Guideline for the Testing of Chemicals. Organisation for Economic Cooperation and Development, Paris, France.

OECD (2010) Short Guidance on the Threshold approach for Acute Fish Toxicity, Series on Testing and Assessment No 126, ENV/JM/TG(2010)/7, OECD, Paris.

OECD (2010). Short Guidance on the Threshold approach for Acute Fish Toxicity – Series on Testing and Assessment No 126, ENV/JM/MONO(2010)17, OECD, Paris.

OECD (2010). Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. 21-day Fish assay, TG 230. OECD, Paris.

OECD (2010). Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. Fish short-term reproduction assay, TG 229. OECD, Paris.

OECD (2011) Guidance Document on the Androgenised female stickleback screen (AFSS). Series on Testing and Assessment No. 148, ENV/JM/MONO(2011)29, OECD, Paris.

OECD (2011) Guidelines for the Testing of Chemicals. Section 2: Effects on Biotic Systems. Test Guideline No 234: Fish Sexual Development Test. OECD, Paris.

Rufli, H. and Springer, T.A. (2011). Can we reduce the number of fish in the OECD acute toxicity test? Environ. Toxicol. Chem. 30: 1006-1011.

Spacie, A. and Hamelink, J.L. (1982). Alternative models for describing the bioconcentration of organics in fish. Environ. Toxicol. Chem. 1: 309-320.

7. POSSIBLE FISH TESTING STRATEGIES

7.1 Introduction

203. The purpose of this chapter is to give some general guidance on possible strategies for approaching hazard testing with fish. There are many ways of tackling this issue, and each jurisdiction will have its own preferences to suit local conditions and policies. No single approach will be "right" in all circumstances. Indeed, given that a testing strategy can and should be influenced by many considerations, including political, economic, and ethical issues, as well as ecological protection goals, it could be argued that no attempt should be made to give over-arching guidance. The intention in this chapter is to illustrate some broad principles which can then be adapted for particular circumstances. It should also be noted that even within a given jurisdiction, different types of chemicals (e.g. pesticides, industrial chemicals, biocides, veterinary medicines) are likely to be subject to different types of testing. Consequently, the strategies suggested in this chapter are only able to illustrate general principles, keeping in mind animal welfare concerns and ensuring the optimal use of available data, rather than specifying particular courses of action.

204. The approaches outlined below do not, of course, operate in isolation. They must be seen in the context of broader requirements for testing other trophic groups in the aquatic environment, the most significant of which at present (in most regulations) are arthropods and algae/plants. In view of the need for performing studies in a cost-effective, yet ecologically protective manner, as well as considering the ethical concerns with the use of fish in toxicity tests, there may sometimes be scope to avoid fish tests altogether. This might be done by relying on the use of non-test methods (e.g. QSAR model predictions), chemical categorization or read across (EC 2003, Bradbury et al. 2004, OECD 2007a), or on tests with invertebrates or other taxa (e.g. see the Species Sensitivity Distribution approach of Posthuma et al. 2001, and the WEB-ICE software developed by the US EPA (http://www.epa.gov/ceampubl/fchain/webice/), although the latter of these approaches have not received wide evaluation or have not been used for regulatory purposes to date. There may also be scope for avoiding excessive use of fish by using limit testing (cf. OECD TG 203) or the threshold approach (OECD GD126, 2010a). There may, furthermore, be scope to use the draft Fish Embryo Test (OECD 2006a), although this test has not yet been fully evaluated. Each of these approaches has limitations which may add an additional degree of uncertainty to conclusions drawn. These uncertainties should be considered before applying these approaches (see Chapter 5).

205. As indicated above, before fish testing is even considered, it is important to question its necessity unless it is required by a particular regulatory authority. This is both because fish tests tend to be more resource-intensive than those with plants and invertebrates, and because of the ethical issues involved in testing vertebrates. Fig. 7.1 presents a generalized flow diagram of a testing framework which can be employed to determine the need for fish testing in a tiered manner depending on the requirements of the regulatory authority.[10] It is acknowledged that various regulatory authorities currently have specific requirements that must be followed. However, the aim of this generic approach is to reflect the latest scientific advances, so that ultimately risk assessment needs will be met and a reduction of vertebrate testing will be realized.

[10]This proposed generic testing strategy makes use of as much prior information as possible in order to determine the need for fish testing. Regulatory requirements in particular jurisdictions may require a more complex or testing-rich assessment. Risk characterization and other chemicals assessments may include, for example, a PBT assessment, hazard classification, and / or endocrine assessment.

7.2 Generic fish testing strategy

206. As illustrated in Fig. 7.1, the problem formulation and analysis phase consists of the gathering of all relevant information, e.g. on physicochemical parameters and environmental fate of the substance, its use pattern and thus potential environmental exposure matrices to consider, and of course any existing information on *in vitro* or *in vivo* activity. Even if fish testing is ultimately regarded as necessary, this initial exercise to gather basic information is essential to guide selection of the most appropriate fish test. These considerations are explicitly discussed *inter alia* by current guidance for the EU REACH legislation.

207. The most fundamental of these data are the physicochemical properties of the test substance, and its fate characteristics. It is impossible to design a reliable fish test without knowing how the substance is likely to behave in the test system. Factors such as water solubility, Henry's Law constant, volatility, octanol-water partition coefficient, and aquatic degradation rate are all needed to decide how the substance should be presented to the fish in order to maintain stable exposure concentrations. Prior knowledge on likely uses of the chemical, including tonnages, use pattern and/or application rates, is another important ingredient in the information needed to underpin fish testing. This is vital for predicting likely exposures (either very approximately in the case of crude tonnage data, or more precisely if expected entry rates to, or concentrations in, the environment are known), and may allow fish testing to be avoided altogether, if surface water exposure is likely to be negligible. It is acknowledged that classification and labelling are hazard-based only and do not require exposure assessment; so this information would not be necessary in those circumstances.

208. If exposure to the aquatic environment is unlikely (e.g. site limited intermediate), then fish testing may not be required. While this is the case for only relatively few substances, it is worthwhile to perform this assessment early in the process to reduce animal testing and to be cost-effective. After this early assessment, it is also important to review any prior testing results, including data from alternative methods, data from analogue chemical(s), or mammalian data on specific mode(s) of action, that might provide relevant information for hazard characterization (e.g. Kaiser et al. 1997, EC 2003). See chapter 5 for more information on alternative methods that might be available to help inform in a weight-of-evidence evaluation in this early stage of the assessment.

209. Some regulatory schemes, although they may differ in concept, bypass the requirement for acute testing of some chemicals (e.g. pharmaceuticals for human use in the European Union and USA (EMEA 2006; FDA-CDER 1998) and require chronic testing in the early stages of risk characterization. The generic framework (Fig. 7.1) allows for moving from gathering data to chronic testing, without the need to generate acute testing results if they are not relevant to the regulatory requirements. Similarly, if, for example, there are indications that the substance would interact with hormonal pathways or there are specific regulatory requirements (e.g. see USEPA Endocrine Disruptor Screening Program –EDSP http://www.epa.gov/endo/), then it could be more relevant to move on to endocrine system-specific testing than to develop standard acute data at an early stage, although the possible need for acute data should not be forgotten. Alternatively, it may be more ecologically relevant to proceed to standard acute and/or chronic testing (i.e. TG 203 or TG 210) after the initial problem formulation and collection of available information.

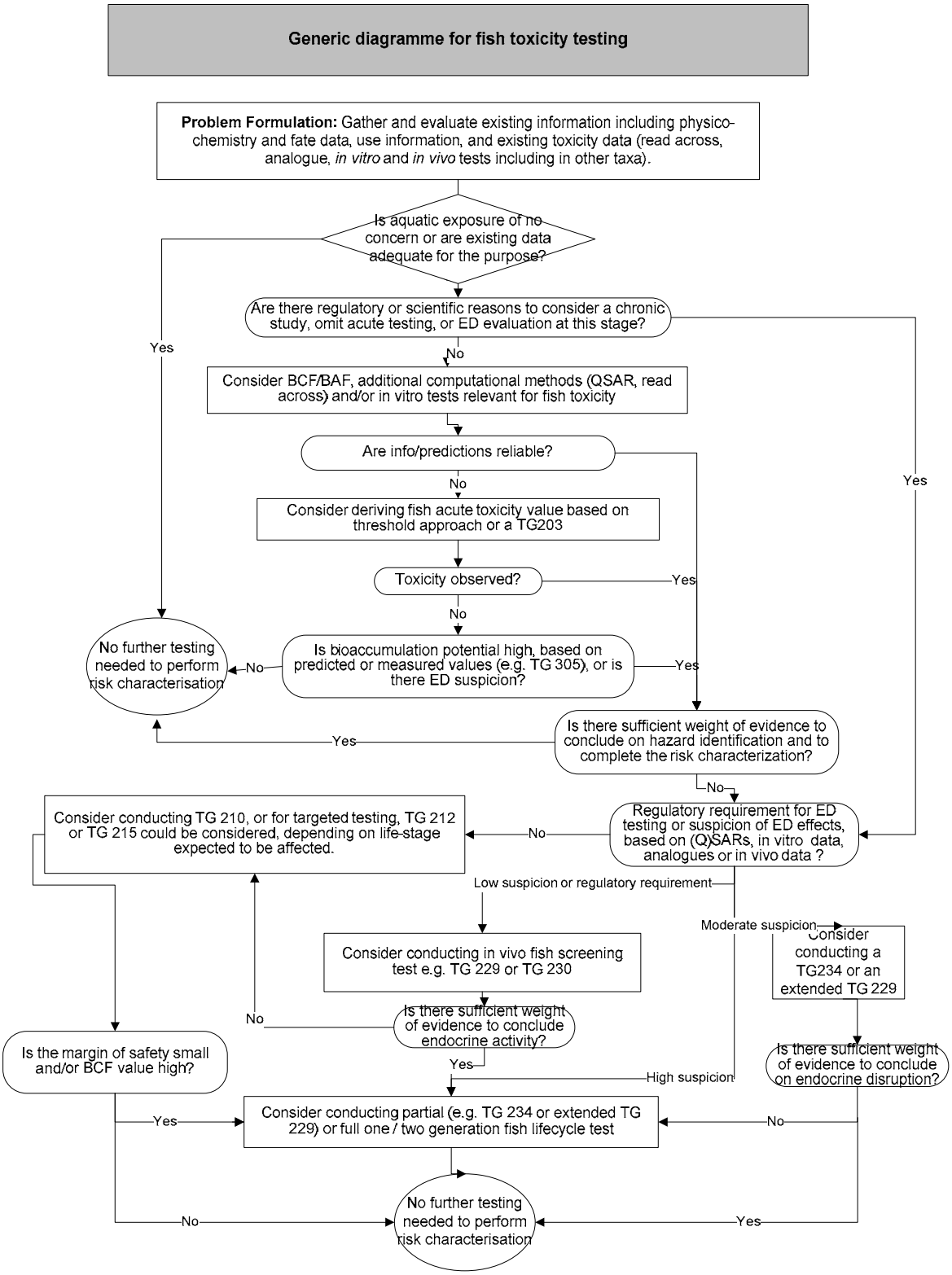


Figure 7.1: A generic testing strategy illustrating general principles of how to optimise fish toxicity testing needs. This strategy makes use of as much prior information as possible in order to determine the need for fish testing. Regulatory requirements in particular jurisdictions may require a more complex or testing-rich assessment. Risk characterization and other chemical assessments may include, for example, a PBT assessment, hazard classification, and / or endocrine assessment.

210. If an early requirement for chronic data is not identified, the need for acute fish toxicity data (OECD TG 203) is now addressed, after first having considered the possibility that existing data can be used to predict fish toxicity using (Q)SARs, read-across or *in vitro* results. The choice of the type of acute testing to perform, or approach to take will depend on the regulatory needs of the geographic region and the regulatory use of the data. For more information on alternative approaches to fish acute toxicity testing, see chapter 5. At this stage, if predictions or other information are reliable, then it may be possible to avoid fish testing and conduct a risk characterization. If scientific justification is provided to show that the substance is likely to be much more toxic in invertebrates or algae/plants, fish toxicity testing may be avoidable (OECD 2010a), depending on the outcome of a consideration of the endocrine activity of the substance. If information is not available or is insufficient to draw that conclusion, or if there are suspicions of possible endocrine activity, then further fish toxicity testing may be needed to define the potential for the substance to interact with the endocrine system. Even if such further fish toxicity testing is not warranted on a substance, an evaluation of its bioaccumulation potential should be performed. If the substance has neither bioaccumulative nor endocrine activity potential, then further testing may not be needed to perform a risk characterization.

211. If the substance is acutely toxic to fish, or if it is predicted or measured to be bioaccumulative and/or potentially endocrine active, then the available weight of evidence for these aspects should be considered. A risk characterization should then be performed and, as appropriate, further testing should be considered. As indicated above, predictions about fish toxicity and/or modes of action may be available from QSAR calculations (e.g. Kaiser et al. 1997, OECD 2006b, OECD 2010b) or read-across. There may also be information from *in vitro* tests with fish cell lines (e.g. vitellogenin (VTG) induction in fish liver cell cultures) or with mammalian cell lines sensitive to particular substances (e.g. the estrogen receptor alpha transcriptional activation assay – OECD TG 455). Furthermore, useful information may sometimes be obtained from toxicity tests with rodents (e.g. OECD TGs 440 or 441), which may be available early during the chemical assessment process, or with other higher vertebrates. This is especially true for endocrine active compounds and analogous substances which operate in vertebrates and act on receptor systems or enzymes that are highly conserved across vertebrate species and classes.

212. It should be noted throughout this proposed fish toxicity testing framework that, if risk characterization indicates an unacceptable risk as determined by a particular regulatory authority, then, as with any risk assessment which is iterative in nature, additional information is likely to be necessary to refine the risk characterization (unless emission reduction is appropriate).

213. If there are no scientific or regulatory reasons for *in vivo* endocrine testing following the risk characterisation, fish early life stage testing (TG 210), or equivalent, should be considered (see further details below). On the other hand, if a targeted endocrine assessment is considered necessary, then the type of testing chosen may depend on the level of suspicion or weight of the evidence for endocrine activity. Also, it should always be borne in mind when considering *in vivo* testing for potential EDCs that it is vital to include the expected most sensitive lifestage as well as the lifestage in which effects are most likely to be manifested. It is, of course, also necessary to ensure that the test has sufficient statistical power for the task in hand, is consistent with the 3Rs, and is optimal in terms of effort and cost (Knacker *et al*. 2010; Schäfers *et al*. 2007). The basis for deciding on the weight of evidence may include positive evidence, negative evidence and lack of evidence, and can be considered in 3 categories, as shown below.

- For example, if only *in vitro* endocrine activity has been observed, or predicted on the basis of read-across or (Q)SARs (i.e. suspicion of endocrine activity is relatively low), OECD TG

229 or 230 may be the most appropriate starting point. Choice between these two screens will be partly driven by whether or not apical information (e.g. fecundity) is needed at this stage in addition to indicators of endocrine activity (e.g. biomarker changes such as VTG). OECD TG 229, for example, informs about interference with fecundity, however, due to the short exposure period (i.e. three weeks), adult life stage and simple test design, it is generally considered a screening test, providing only qualitative apical information for that endpoint.

- If, based on the available data there is already a moderate suspicion for the substance being endocrine active, e.g. if *in vivo* endocrine-related effects have already been observed in another vertebrate taxon (e.g. OECD TG 440 or TG 441), or if persuasive read-across or QSAR predictions or *in vitro* data are available, then a Fish Sexual Development Test (TG 234) or an Extended TG 229[11] could be considered because the mode of action and therefore the likely most sensitive lifestage should be known. Note that if either of these tests are conducted, early life stage tests such as TG 210 (if triggered at any point) may not be necessary to perform a risk characterisation since the endpoints of TG 210 are adequately covered by these two tests.

- If, in addition to the above mentioned types of available data, positive *in vivo* data are also available from other vertebrate taxa or higher tier vertebrate tests (i.e. suspicion of endocrine activity and disruption is high) and if establishment of a NOEC for adverse effects in fish caused by endocrine disruption is essential in the regulatory context, then a full life-cycle fish test (FFLCT), or a medaka multi-generation test (MMGT) or similar multigeneration test including endocrine disrupter- related effect endpoints beyond more traditional adverse effect endpoints, may be warranted without intermediate steps. The full life-cycle and the two-generation tests (which include relevant endocrine-related effects and which are currently in development) integrate effects across all life stages. Alternatively, partial life-cycle tests such as the Fish Sexual Development Test (FSDT – OECD TG 234) can also inform about endocrine-system related interference with sexual development (sex ratio and secondary sex characteristics in some species besides information about VTG production in fish), among other endpoints. Similarly, when an extended TG 229 (another partial lifecycle test) is adopted by OECD, that could sufficiently inform about endocrine-related interference with fish reproductive parameters. If mode of action information is available, then choice of appropriate test method (i.e. partial or full lifecycle tests) may be assisted (ECETOC, 2007; Knacker *et al* 2010). Consultation with regulatory authorities concerning choice of the most appropriate test method is also recommended.

214. If screening with TG 229 or TG 230 is chosen, positive results in either screen would trigger the need to conduct either a FSDT (TG 234), a partial life-cycle test focusing on reproduction (e.g. an extended OECD TG 229), or an FFLCT or MMGT. If the screening assays are negative or when there is no suspicion for endocrine activity substantiated by e.g. negative *in vitro* and *in silico* data, absence of endocrine-related effects in other taxa and/or by reading across from other comparable structurally-related chemicals, the initial test which should be considered if fish chronic toxicity data are required is the early life-stage test (OECD TG 210). In particular circumstances where juvenile growth is likely to be susceptible, the juvenile growth test (OECD TG 215) may be preferred by some regulatory authorities. The relatively brief egg and sac-fry test (OECD TG 212)

[11] An extended version of TG 229, the development of which is under discussion, could more accurately be described as a Fish Partial Lifecycle Reproduction Test, but for reasons of brevity is described in this document as an 'extended TG 229'. It would essentially consist of a combination of TG 229 and TG 210, thus covering the part of the fish lifecycle involving reproduction and early development, but not that involving sexual differentiation. An example of this test in action is given by Panter *et al*., (2010). It is seen as being complementary to the FSDT (TG 234) and might be used if reproduction rather than sexual development was expected to be the most sensitive part of the lifecycle.

may be considered if effects on larval development are expected to result from short-term exposures alone. However, it appears that in reality, testing for regulatory purposes using either OECD TG 212 or OECD TG 215 is generally infrequent. With a shorter exposure time and a single life-stage target, the embryo and sac-fry test could be less sensitive than the full early life-stage (ELS) test (OECD TG 210), particularly with respect to chemicals with high hydrophobicity (log K_{OW} > 4) and chemicals with a specific mode of toxic action (OECD 1998). However, smaller differences in sensitivity between the two tests would be expected for chemicals with a non-specific, narcotic mode of chronic toxic action (Kristensen 1990).

215. At the end of this generic testing strategy, the outcome of the early life-stage test (TG 210) may be sufficient to complete the risk characterization. However, where the margin of safety is small (i.e. toxicity is high by comparison with predicted exposure), where there is a high bioaccumulation potential, and/or if the substance is likely to cause prolonged exposure, it should be considered whether to conduct a full or multiple life-cycle study so as to base the predicted no-effect level on the most sensitive endpoint.

216. It is apparent that skilled scientific judgement, along with a sound knowledge of regulatory requirements, is needed to make many decisions in this generic strategy. Choices are rarely clear-cut, and must make efficient use of all available information on the tested substance. This will often involve using a weight-of-evidence approach, a subject described in more detail elsewhere (e.g. see OECD 2011b, or the REACH Endpoint Specific Guidance). Each regulatory authority may wish to set different "triggers" for progressing further in the testing framework (e.g. the precise values of log K_{OW} which might trigger a fish bioconcentration test; or the size of a safety margin that might obviate the need for longer-term testing, etc.).

7.3 Influence of exposure type on testing strategy

217. The type of exposure expected to be caused by a substance will have a profound influence on the effects it is likely to cause in fish, and hence on the type of fish screening or testing which it is most efficient or relevant to conduct. Below, considerations are given with respect to short- and long-term exposure, pulsed exposure and exposure *via* water, food and sediment.

7.3.1 Shorter-term exposure toxicity tests

218. With the exception of substances (such as EDCs or strongly bioaccumulative chemicals), which may have the potential to cause long-term effects from even very brief exposures (sometimes just at sensitive life-stages), expected short-term exposure (i.e. a few days) will usually be a trigger for short-term testing (e.g. OECD TGs 203, 204, or potentially the fish embryo test). In this context, however, it should be remembered that even substances that disperse or degrade rapidly and, thus, may be thought of as only likely to exert short-term exposure, may nevertheless be subject to continuous or semi-continuous discharge to the waters which fish inhabit. Such substances have been termed "pseudo-persistent", and short-term testing alone may not be sufficient to characterise their environmental hazards (Fent et al., 2006). Equally, if the acute fish toxicity level is close to the foreseeable acute exposure level, this may also be an indication that longer-term testing is desirable. It would be useful to conduct a review of the sensitivity of different fish life-stages to different groups of chemicals in order to help support decisions about which type of acute test may be the most appropriate in particular circumstances.

219. If a substance causing, or expected to be causing, short-term aquatic exposure is suspected to interact with the endocrine system on the basis of prior *in silico*, *in vitro* or *in vivo* data, a short-term *in vivo* screen that is sensitive to some endocrine active compounds (e.g. OECD TG 229 or

230) may be desirable. These screens can provide information about the likelihood that short-term exposure at the sexually-mature stage could cause longer-term effects with implications for reproductive output and population stability, although they cannot provide reliable information for risk assessment. However, they may be insensitive to anti-androgens, in which case the androgenised female stickleback screen (OECD GD 148) should be considered. These screens are also insensitive to thyroid-active substances, and to EDCs which may act on the corticosteroid system etc. However substances interfering with the thyroid hormone system may in some cases be identified by use of non-test information and/or data from the amphibian metamorphosis assay or mammalian repeated dose toxicity studies in rats.

7.3.2 Longer-term exposure toxicity tests

220. If significant exposure is expected to last longer than a few days, then longer-term testing will usually be desirable. A range of sub-chronic tests with fish is available, including OECD TGs 210 and 215, although several useful fish partial and full life-cycle tests are currently available (TG 234) or undergoing validation and are expected to be developed into OECD guidelines in due course (e.g. FFLCT, OECD 2008a; MMGT, OECD 2002). Also, for substances which are expected to cause longer-term exposure, or are expected to bioaccumulate (e.g. high octanol-water partition coefficient), the fish bioaccumulation flow-through test (OECD TG 305) can provide useful confirmatory data. A revised version of OECD TG 305 allowing dietary exposure for hydrophobic substances is currently under development.

221. Choice of which longer-term toxicity test to conduct will be driven by several considerations including the suspected mode of action. Thus, for example, substances expected to cause non-specific systemic toxicity may only need to be tested in OECD TG 210, which is generally regarded as providing good predictivity for effects to be expected over a whole life-cycle (McKim, 1977). On the other hand, if highly specific modes of action (such as the inhibition of a key enzyme or interaction with a specific hormone receptor) are suspected or known to be operating, and particularly if there is expected to be a window of high sensitivity during a part of the life-cycle not covered by OECD TG 210 (e.g. sexual development or reproduction), then there may be no option but to conduct a partial life-cycle test (e.g. Fish Sexual Development Test TG 234, or an extended TG 229 including early life-stages) or full life-cycle test which covers all sensitive stages.

222. For endocrine active compounds which specifically target sexual development, it would be possible to conduct a FSDT (TG 234) and probably obtain good predictivity about certain types of adverse endocrine-related effects that could occur in a life-cycle test. Currently available, but limited information seems to suggest that a positive result under such circumstances might indicate a need for full life-cycle testing (e.g. FFLCT or MMGT), if the positive test is not considered already sufficient for a definitive hazard or risk assessment. A negative FSDT may imply that further endocrine-related fish testing is not highly warranted unless the weight of evidence suggests otherwise. A positive result in the fish short-term reproductive screening assay (OECD TG 229) could be indicative of effects during the reproductive part of the life-cycle, but as exposure in this test is relatively brief (21 d) and does not occur during development of the fish, negative results might also need to be followed up with the FSDT (TG 234) if available information from other sources suggests some evidence of endocrine activity.

223. Depending on regulatory requirements, if short- or long-term exposures occur which may cause adverse effects over sensitive parts or the whole span of a fish life-cycle, a decision would have to be made about whether an FFLCT (e.g. US EPA 850.1500; Benoit 1982; Länge et al 2001; OECD, 2008a) or the MMGT (or similar multigeneration test) are more appropriate. More research is required on this point, but it seems likely that many substances (including some endocrine active

compounds) will show similar potency in the two types of life-cycle test (OECD 2008a). However strongly bioaccumulative substances may cause impacts in the subsequent generation *via* maternal transfer of residues to eggs.

224. Finally, based on studies conducted in mammalian species, it is also possible that epigenetic effects would only appear in later generations (Vandegehuchte and Janssen, 2011). For such substances, their possible implications for aquatic life should be considered. There are on-going efforts to address these effects (see OECD 2011a).

7.3.3 Pulsed exposure

225. Pulsed exposure of aquatic organisms, e.g. to pesticides in surface water, is often a result of application to crops several times in succession. Semi-continuous discharges of industrial chemicals may also result in pulsed exposures. Depending on the characteristics of the substance, and on the size and frequency of pulses, effects can be produced in long-term tests which are similar to those resulting from continuous exposure. Consequently, the advice in section 7.3.2 concerning long-term exposures also usually applies to pulsed exposures. However, in some circumstances, it would be appropriate to run the standard long-term fish tests, but arranging for exposures to be pulsed in a way which simulates expected exposures in the environment even though such exposure regimes are not a normal standard procedure in regulatory testing requirements. Regulatory authorities should be consulted before considering these exposure scenarios.

7.3.4 Exposure via water (see also the chapter 4 on general test considerations)

226. All fish-related OECD TGs and draft TGs are primarily designed for testing exposure *via* the water phase. At least for weakly- or non-bioaccumulated substances, or for substances that do not sorb strongly to sedimentary particles, this may be the most realistic route of exposure in fish. Guidance for dosing *via* the water phase is generally given in OECD TGs and in the OECD Guidance Document on aquatic toxicity testing of difficult substances (OECD GD No. 23), but it is important to bear in mind that rapidly degraded, rapidly volatilized, or strongly sorbed substances may disappear quickly from test waters, and therefore adequate analysis of such waters is warranted. These properties will also drive the choice between static, static-replacement, and flow-through fish tests. In addition, these properties will influence decisions about whether to use solubilising agents such as organic solvents, or instead to use methods such as saturated desorption columns etc. The possible presence of important degradation products should also be considered as testing is planned. Finally, in tests where fish are exposed *via* the water phase, excessive fish loading or strong bioconcentration may rapidly deplete target exposure concentrations. This problem will rapidly become apparent by monitoring measured exposure concentrations.

7.3.5 Exposure via food or sediment

227. Substances that are insoluble in water, are strongly adsorptive to the test vessel, and/or are very lipophilic are almost impossible to test adequately in fish when dosed into the test water. In cases of this type, or if the concern is primarily for substances to which wild fish are mainly exposed *via* sediment or food, it may be more appropriate to consider exposure *via* these alternative matrices. OECD TGs are not primarily designed for this approach, but there is scope for adapting them so that fish either come into contact with contaminated sediment during the test, or are fed with food that has been dosed with test substance. Both of these dosing methods are not without their problems, particularly in relation to spiking procedures for food items or sediment, and calculation of the precise dose received. Spiked food could also release the substance into the water phase thus changing the type and concentration of exposure in the test. Since normal hazard and risk assessment

procedures are based on effect concentrations in the test water and not on dietary doses, the regulatory use of such data should be identified prior to testing.

7.4 Interpretation and conclusions

228. Some guidance has already been given earlier in this document on the interpretation of fish test data, and many TGs go into limited detail on this subject. In addition, an OECD guidance document on the interpretation of tests for endocrine active compounds is in preparation (OECD, 2011b). Probably the most important point concerning interpretation is that test data should never be considered in isolation, but should be evaluated with all other relevant data by experienced scientists using a weight-of-evidence approach.

229. Perhaps the most difficult situations arise when a test meets quality/performance criteria, but the results nevertheless seem equivocal, or they appear to conflict with data from another test. The first situation may include, *inter alia*, non-monotonic concentration-responses, responses that just fail to be statistically significant, examples of receptor-mediated toxicity partially masked by systemic toxicity, and (in higher tier tests) changes in biomarkers of effect without accompanying apical impacts at the same exposure level or within the time frame of the study. Data of this type should not necessarily be ignored – they may still be revealing something of value. For example, some endocrine active compounds may genuinely give non-monotonic responses, and effects just below the level of statistical significance may simply mean that the test substance is very weakly-acting. Conversely, some apparently receptor-mediated effects (e.g. vitellogenin depression in female fish) may in fact be caused by systemic toxicity, while others may be genuine responses to an endocrine active compound that have been partially counteracted by systemic effects. Careful attention to the dosing regime including aspects of adsorption, distribution, metabolism, and excretion (ADME) may be needed to disentangle such problems. Finally, biomarker responses (e.g. VTG induction in males in the FSDT) may possibly occur without corresponding alterations in apical endpoints (e.g. sex ratio in the FSDT). This may simply indicate that the biomarker is more sensitive to the underlying cause (i.e. estrogen exposure) than the apical endpoint, but it could also just imply that the test was too short to record effects on the apical endpoint. Another example may be that if the response variable sex ratio changes it should be regarded as an endocrine mediated effect in the absence of counter-evidence. When exposure is to an androgen such as trenbolone causing sex ratio change, the observed lack of VTG response in the test is easy to explain: VTG induction in males is not to be expected, and VTG depression in females is often not possible to observe because of the potency of the substance and the design of the FSDT. In any case, such apparent conflicts or peculiar findings will provide useful information when deciding if more advanced testing is desirable.

230. As a general rule in the second situation, results from a higher tier test should be considered to trump or supersede those of a lower tier. So, for example, if a fish (full) life-cycle test reveals adverse effects, while a previous partial life-cycle test did not, the full life-cycle test should be used in preference for risk assessment. A note of caution is, however, warranted here: every time apparent conflicts between tests are observed (and this is also the case when results of tests from different tiers of the OECD EDTA Conceptual Framework are being compared), it is important to evaluate what the cause may be. Have the same endpoints for example been addressed in an equally sensitive way? Are the sensitivity of the species and response variables of the endpoints used comparable? When test data on a given tier (e.g. screening tests for endocrine activity) appear to conflict, it may be sensible to be cautious and use the positive response as a basis for movement within the test strategy. In such situations, however, it could also be desirable to seek confirmatory data from repeat tests.

231. The most comprehensive type of fish testing covers at least one full life-cycle, and greatest reliance can generally be placed on tests of this type. When faced with lower tier data (e.g. from the early life stage test, OECD TG 210), the first question should be whether the outcome provides enough information for a regulatory decision to be made. In some cases, further generation of information from additional fish testing might still be warranted. Scientific experience will be necessary to decide under which circumstances further testing can be safely avoided, employing a weight-of-evidence approach. This is not the place to provide detailed guidance on the interpretation of fish life-cycle test data, but it is clear that the distinction (described above) between biomarker and apical endpoints is of relevance to these tests, even though it is realised that some endpoints may contain information which may be considered both apical and biomarker related. While biomarkers may provide mechanistic data about a possible endocrine active compound, only the apical endpoints relating to adverse effects can be used directly in definitive environmental hazard and risk assessment, which seeks to protect fish populations and other aquatic species. In other words, while both mechanistic and apical data are needed to categorize a substance as an endocrine disrupter, mechanistic data alone can raise suspicion, and apical data alone are sufficient for traditional hazard and risk assessment. Other relevant questions which may be important to consider before further fish toxicity testing is concluded include: What is the sensitivity or power of the test considered? What type of information does the test deliver as regards "adverse effects" and modes of action? Is such information useful for hazard and risk assessment or only for triggering further testing? These considerations should focus on the most efficient testing strategy (in terms of cost and animal usage) taking account of current knowledge.

232. Finally, when following a fish hazard assessment strategy of whatever type, it is essential to keep in mind the objectives of testing. Criteria for hazard classification will generally be clear and unambiguous, but if it is intended to use the data for categorizing a chemical as an endocrine disrupting compound or for use in hazard categorisation or risk assessment (where effects and exposure levels are compared), there is often a temptation to seek just one more piece of information (on the potentially spurious grounds that more equals better). Such a temptation should be resisted, and it must be remembered at all times that the primary objective of fish testing is to protect the stability of populations of fish and other pelagic species and to safeguard consumers of fish (humans and wildlife).

7.5 References

Benoit, D.A. (1982) User's guide for conducting life-cycle chronic toxicity tests with fathead minnows (*Pimephales promelas*). *Environ. Res. Lab*. – Duluth, MN. EPA 600/8-81-011.

Bradbury, S., Feijtel, T., van Leeuwen, K. (2004) Meeting the scientific needs of ecological risk assessment in a regulatory context. Environ. Sci. Technol. 38:463a-470a.

EC (2003) Technical guidance document in support of Commission Directive 93/67/EEC on risk assessment for new notified substances, Commission Regulation (EC) no. 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. European Chemicals Bureau, Joint Research Centre, Ispra (VA), Italy.

ECETOC (2007). Intelligent testing strategies in ecotoxicology: mode of action approach for specifically acting chemicals. ECETOC Tech. Rept. 102, European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels, 145 pp.

EMEA (2006) Guideline on the environmental risk assessment of medicinal products for human use. EMEA/CHMP/SWP/4447/00 corr 1, European Agency for the Evaluation of Medicinal Products, London

FDA-CDER (1998).Guidance for industry – environmental assessment of human drugs and biologics applications. Revision 1. FDA Center for Drug Evaluation and Research. Rockville, Arkansas.

Fent, K, Weston, A.A. and Caminada, D. (2006). Ecotoxicology of human pharmaceuticals. Aquat. Toxicol. 76, 122-159.

Kaiser, K.L.E., Niculescu, S.P., McKinnon, M.B. (1997) On simple linear regression, multiple linear regression, and elementary probabilistic neural network with Gaussian Kernal's performance in modeling toxicity values to fathead minnow based on Microtox data, octanol/water partition coefficient, and various structural descriptors for a 419-compound dataset. In: Chen, F. and Schüürman, G. (eds) Quantitative Structure-Activity Relationships in Environmental Sciences – VII. Society of Environmental Toxicology and Chemistry, Pensacola, Florida. pp. 285-297.

Knacker, T., Boettcher, M., Rufli, H., Frische, T., Stolzenberg, H.C., Teigeler, M., Zok, S., Braunbeck, T. and Schäfers, C. (2010). Environmental effect assessment for sexual-endocrine disrupting chemicals – fish testing strategy. Integr. Environ. Assess. Manag. 6: 653-662.

Kristensen, P. (1990) Evaluation of the Sensitivity of Short Term Fish Early Life Stage tests in relation to other FELS test methods. Final Report to the Commission of the European Communities, 60 pp.

Länge, R., Hutchinson, T.H., Croudace, C.P., Siegmund, F., Schweinfurth, H., Hampe, P., Panter,G.H. and Sumpter, J.P. (2001). Effects of the synthetic estrogen 17-alpha-ethinylstradiol on the life-cycle of the fathead minnow (*Pimephales promelas*). Environ. Toxicol. Chem. 20: 1216-1227.

McKim, J. (1977) Evaluation of tests with early life stages of fish for predicting long-term toxicity. J. Fish. Res. Board Can. 34:1148-1154.

OECD (1998) Fish short-term toxicity test on embryo and sac-fry stages, TG 212. Organisation for Economic Cooperation and Development, Paris. 20 pp.

OECD (2002) Fish two-generation test guideline. Draft proposal for a new guideline. Organisation for Economic Cooperation and Development, Paris. 18 pp.

OECD (2006a) Fish embryo toxicity (FET) test. Draft proposal for a new guideline. Organisation for Economic Cooperation and Development, Paris. 11 pp.

OECD (2006b) Report on the regulatory uses and application in OECD member countries of (Quantitative) Structure-Activity Relationships [(Q)SAR] models in the assessment of new and existing chemicals. Series on Testing and Assessment no. 58, Organisation for Economic Cooperation and Development, Paris. 79 pp.

OECD (2007a) Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. GD 69. Organisation for Economic Cooperation and Development, Paris, 154 pp.

OECD (2007b) Phase 1 of the validation of the fish sexual development test for the detection of endocrine active substances. Organisation for Economic Cooperation and Development, Paris. 67 pp.

OECD (2008a) Detailed review paper on fish life-cycle toxicity testing. OECD Series on Testing and Assessment No. 95, Organisation for Economic Cooperation and Development, Paris. 162 pp.

OECD (2008b) Detailed review paper on the use of metabolizing systems for in vitro testing of endocrine disrupters. OECD Series on Testing and Assessment No. 97, Organisation for Economic Cooperation and Development, Paris. 95 pp.

OECD (2010a), Short Guidance on the Threshold approach for Acute Fish Toxicity , Series on Testing and Assessment No. 126, ENV/JM/TG(2010)/7, OECD, Paris.

OECD (2010b) QSAR Application Toolbox, Version 1.1.02. Organisation for Economic Cooperation and Development, Paris. 5 pp.

OECD (2010c) Fish Sexual Development Test - Draft proposal. Unpublished document, Organisation for Economic Cooperation and Development, Paris. 21 pp.

OECD (2011a). Draft detailed review paper on the state of the science on novel *in vitro* and *in vivo* screening and testing methods and endpoints for evaluating endocrine disruptors. Chapter on 'Endocrine disruptors and the epigenome', OECD, Paris. http://www.oecd.org/dataoecd/42/53/48435503.pdf.

OECD (2011b). Draft Guidance Document on standardised test guidelines for evaluating chemicals for endocrine disruption, v.12, Aug. 2011. http://www.oecd.org/document/12/0,3746,en_2649_34377_1898188_1_1_1_1,00.html

Panter, G.H., Hutchinson, T.H., Hurd, K.S., Bamforth, J., Stanley, R.D., Wheeler, J.R. and Tyler, C.R. (2010). Effects of a weak oestrogenic active chemical (4-tert-pentylphenol) on pair-breeding and F1 development in the fathead minnow (*Pimephales promelas*). Aquat. Toxicol. 97, 314-323.

Posthuma, L., Suter II, G.W., Traas, T.P. (2001) Species sensitivity distribution in ecotoxicology. Lewis Publ., 276 pp.

Schäfers, C., Teigeler, M., Wenzel, A., Maack, G., Fenske, M. and Segner, H. (2007). Concentration- and time-dependent effects of the synthetic estrogen, 17α-ethynylestradiol, on reproductive capabilities of the zebrafish, *Danio rerio*. J Toxicol Environ Health, Part A, 70: 768-779.

Vandegehuchte, M.B. and Janssen, C.R. (2011). Epigenetics and its implications for ecotoxicology. Ecotoxicol. 20: 607-624.

ANNEX

Considerations and recommendations

(As agreed by the Workshop held on 28-30 September 2010, at Jealott´s Hill, United Kingdom)

The present chapter summarises the discussions of the Workshop on a Fish Toxicity Testing Framework held from September 28 to 30, 2010, at Jealott´s Hill, UK, and presents the conclusions and recommendations of the workshop.

1 Is there a need for consistency in the definitions of fish life-stages among test guidelines?

Recommendations:

- There is a clear need for harmonization of the definitions across aquatic Test Guidelines. The formation of an expert group is recommended.
- As a deliverable, a guidance document should harmonise the use of commonly found terms and facilitate the revision of the test guidelines. These terms should not only cover life-stages (e.g. F0, F1), but also terms such as acute, chronic, sub-chronic, death, moribund, sublethal, spawning status, etc.

2 Should biomarker endpoints be included in fish lifecycle tests?

Considerations:

- The regulatory context is important and should always be considered.
  - If, for example, the mode or mechanism of action (MOA) is the basis for regulation, then the use of biomarkers is more important than in regulations based on apical responses.
- Fish full life-cycle (FFLC) studies may include biomarkers, and their inclusion should be guided by the suspected mode of action, employing a weight-of-evidence approach. E.g., if good mechanistic data are already available, re-assessment of biomarkers might not be required, although they might be helpful in confirming the cause of certain apical responses.
- In the context of the "3Rs", collecting biomarker information and any additional information may be a useful approach.
  - Inclusion of biomarkers should be guided by the suspected mode of action, employing a weight-of-evidence approach.
  - Biomarkers might also be helpful in gaining a better understanding of specific mode of action that can aid in the development of alternative testing strategies.

3 How should potential solvent effects be handled?

Recommendations:

- An international workshop should be organised to pull together available information on solvent and dispersant use in testing of difficult substances, their effects on control animals, and the statistical analysis of solvent and dispersant controls, with the aim of updating guidance on difficult-to-test substances and, where needed, accounting for non-solvent technologies.

  The workshop should address two major issues:

  - Statistical procedures: How to handle the analysis of studies using solvents? Analysis when, and if, both solvent and water controls are needed.
  - Exploration of existing information: What is known about the interaction between a test substance and solvents/dispersants? How can such interactions be recognized and dealt with?

- This could be followed by an update of the TGs; (1) to replace the individual recommendations on use of solvents/dispersants by referring to the updated guidance on difficult-to-test substances, and (2) to reduce animal use by the removal of the dilution-water control, where possible.

4 Is it appropriate to determine EC_x and NOECs from the same study design?

Considerations:

- The NOEC should have sufficient power to detect biologically relevant effects.
- EC_x may be beyond the range of control variability.
- Ideally, in the EC_x approach, x should be between two tested concentrations. It is important to recognise that extrapolation beyond the range of data adds "significant" uncertainty and needs to be justified.

Conclusions:

- Appropriateness of determining EC_x and NOECs from the same study design depends on the data set and test design.
- For certain existing regulatory frameworks, it might be appropriate to focus on NOEC test designs for fish chronic endpoints (e.g. Federal Insecticide Fungicide Rodenticide Act, FIFRA).
- For future regulatory frameworks, it could be required to have both EC_x and NOEC determinations in fish chronic studies (e.g., Directorate General for Health and Consumer Protection, DG Sanco, 2010). However, this has serious implications for experimental design, time and cost, ethical and statistical interpretation. It might not be practical to design tests with multiple endpoints to determine both the NOEC and EC_x values for endpoints of interest.

5 How should test concentrations be chosen for fish endocrine screening assays (OECD TGs 229 and 230)?

Considerations:

- There are difficulties with setting concentrations without unnecessary use of animals, including the possible need to repeat a study due to excessive general toxicity and observed mortality.

- Approaches will vary depending on the availability of pre-existing data, although a general approach might consider:

  - range-finding for the determination of the LC_{50} (considering water solubility, maximum test concentration of 100 mg/L; range-finding design as discussed in chapter 3);

  - standard 96h LC_{50} (colour, behaviour, mortality, gross morphology) to provide information to determine the maximum tolerated concentration (MTC) or similar approaches;

  - estimate MTC or similar approaches for screening assays or fish chronic test data.

Recommendations:

- The Fish Drafting Group should provide guidance on how to determine the most appropriate concentration range including evaluation of existing data and appropriate range-finding.
  - As a possible starting point, Hutchinson et al. (2009) might be considered.

References

Hutchinson, T.H., Bögi, C., Winter, M.J., Owens, J.W. (2009) Benefits of the maximum tolerated dose (MTD) and maximum tolerated concentration (MTC) concept in aquatic toxicology. Aquat. Toxicol. 91: 197-202.

6 Is there a need for guidance on the interpretation of acceptance/validity criteria?

Considerations:

- Minor statistical deviations from acceptance/validity criteria should not be used to reject scientifically sound studies. The consequences of these deviations should be explained. The use of historical control data and knowledge of species sensitivity towards test conditions (e.g. dissolved oxygen, temperature) could be used to support the argumentation.

- There is a need to acknowledge that the longer the study, the more challenging it is to meet the acceptance/validity criteria.

  - Multiple acceptance/validity criteria increase the chance that a study is judged invalid (probability theory) and this should be recognized in the context of a weight-of-evidence decision regarding overall test acceptability.

Recommendations:

- Acceptance criteria should be handled holistically: Is it just one criterion or all criteria failing?

- In order to review the different validity criteria in fish toxicity TGs, it is recommended to contact contract laboratories and regulatory authorities to determine which criteria often fail – also recording food and water quality in these studies (see section 8.7. below) – to come up with recommendations for improvements.
- Frequent failure of criteria by established laboratories could be used to drive re-assessment of the criteria levels.
- A mechanism should be developed to collect information on newly deployed test methods to decide whether stated validity criteria are realistic or need adjustment. Guidance should be developed on how to meet acceptance/validity criteria.
- A table should be created to quantify the potential for failing one or more validity criteria by chance.

7 Can we develop guidance on ensuring consistent and acceptable water and nutrition quality?

Considerations:

- Review approaches (literature and existing guidance documents) used by laboratories and other testing facilities to assess levels of contaminants (including relevant pathway-based assessments, such as *in vitro* bioassays) that may be present in water and feed (exercise linked to acceptance/validity criteria).
- Promote good practice for food quality monitoring.
- Any modifications need to remain technically feasible.
- Guidance to reduce variability in responses should be provided as a refinement approach
  - *Additional criteria should be developed based on the purpose of the test (e.g. EDCs).*
- Water quality criteria should be defined (example contained in OECD TG 215).

Recommendations:

- A survey of laboratories should be carried out with wide coverage to capture best practice regarding water and nutrition quality, including the presence of contaminants of concern in water and feed. This survey could be linked to the mechanism for collecting information on newly deployed test methods recommended under the previous point (section 8.6: exercise on acceptance/validity criteria).

8 How can reduction of animal use be achieved without compromising the statistical power of the test?

Considerations

- Whenever a TG is revised, it would be appropriate to investigate the statistical power of endpoints measured, and any possibility to reduce animal use without reducing statistical power should be analysed.

- Given the potential failure of a test conducted with too few animals, the appropriate number of animals and replicates should clearly be defined on the basis of statistical analyses.

- It is just as unethical to use too few animals as too many, if the results of such a test cannot be used. If a guideline revision is considered with the aim of reducing the number of animals, this statistical re-evaluation (including investigation of statistical power) should be part of the package.

    - There have been retrospective analyses done for OECD TGs 203 and 210.

    - For alternative methodologies, there is also a need for a statistics-based comparison of the new and existing approaches in order to determine if the alternative represents an improvement or not. This is needed to determine whether the data are sufficient to address regulatory needs and are acceptable for MAD.

    - Reducing the number of animals is achieved on a biological response basis by examining the trade-offs between the number of fish per replicate and the number of replicates per test concentrations, and the selection of the statistical test methods/model to be fitted.

    - In some cases, gains could be made by redesigning the TGs. Generally making these types of changes requires extensive study.

    - For this reason, it is difficult to design optimised tests for both NOEC and EC_x values based on the same biological endpoints within the same study.

9 Is there a need for replication in OECD TG 203?

Conclusion:

- OECD TG 203 is a very robust study; this test has been used for many years with only rare indication of more non-monotonicity in the concentration response than simple binomial probability would predict.

- With replication, i.e., additional tanks in each test concentration, it would be possible to test for extra-binomial variance. If no evidence is found of extra-binomial variance through a sufficiently powerful test, then the sample percents can be regarded as unbiased estimates and probit analysis or other regression techniques can be used to estimate the LC_{50}.

- Without more replication (three tanks per test concentration would be ideal), there is no theoretically sound way to evaluate whether the sample percents are appropriate for the purpose of estimating an LC_{50}.

- While evidence of a monotonic concentration-response is not unassailable, there still seems insufficient justification to increase the replication from one to three tanks per test concentration and control.

10 What are the potential criteria for considering deletion or update of a guideline?

Considerations:

- Potential criteria for deletion of a TG:

    - no current use;

- better alternative guideline available;
- ethical indefensibility;
- no longer scientifically/biologically valid.

- Potential criteria for update of a TG:
  - Optimisation of sensitivity (i.e. availability of more accurate tests or endpoints covered by another TG);
  - Improvement of scientific defensibility;
  - Enhancement of ethical defensibility;
  - potential for increased applicability after revision (e.g. improved acceptance of TG 212 after addition of appropriate feeding);
  - increased cost-benefit.

11 Recommendations for deletions of OECD TGs

11.1 Deletion of OECD TG 204

Considerations:

- rarely used (OECD survey 2009, see Annex);
- extension of OECD TGs 203 or 215 considered as better alternatives;
- ethical indefensibility: relevance of OECD TG 204 questionable;
- no longer scientifically valid for use as a chronic study.

Recommendation:

- Deletion of TG 204.

11.2 Deletion of TG 212

Considerations:

- rarely used, though some member countries may find this sub-chronic test a candidate protocol for their future implementation of effluent regulation;
- alternative guideline is potentially available (i.e. Fish Embryo Toxicity test, if validated);
- lack of feeding could be considered as ethically indefensible;
- no longer scientifically valid for the following reasons:
  - recommended time to start feeding too late,
  - TG considered relatively insensitive,

- cannot be considered as a chronic test, but may be suitable as a sub-chronic test or range-finder for some countries.

Recommendation:

- Consider action following the completion of FET test validation, e.g. deletion or modification of TG 212.

12 Recommendations for modification of OECD TGs

12.1 Modification of OECD TG 215

Considerations:

- OECD TG 215 may have some uses (e.g. pesticide intermediates); OECD TG 215 is the preferred methodology for feeding studies;
- The Early Life-Stage test (OECD TG 210) could be considered as an alternative guideline; however, OECD TG 210 uses more fish, and, for certain compounds, OECD TG 215 has been found to be more sensitive than the ELS test (not published);
- ethical indefensibility: not relevant for this TG;
- no longer scientifically valid: not relevant for this TG.

Recommendation:

- No deletion.
- Update may be warranted, possibly concerning clarity about replication.

12.2 Modification of OECD TG 210

Recommendation based on analyses by J. Oris (manuscript in preparation):

- Recommendation to raise the allowable hatch/survival performance criteria (potentially with different percentages depending on the species).
  - Modification would allow the use of fewer animals without compromising statistical power.
- Replicate should be clearly defined as the test vessel because of binomial over-dispersion.
  - Increase the number of replicates, but reduce the number of animals per replicate: Based on an analysis of control treatment performance, the number of replicates should be increased from a minimum 2 to 4 whilst using the same number or fewer animals due to the definition of a replicate.
- Guidance on wet weights *versus* dry weights should be considered (based on this analysis, wet weights appear to be less variable and therefore more sensitive than dry weights).
- There should be closer consideration of parameters that might affect the sensitivity of various endpoints, e.g. tank size, temperature, oxygen, water quality, etc.

- There needs to be additional explanation /description of which length measurements should be used.
- There is a need for consistent guidance on thinning after hatch, since this can affect growth rates.
- Feeding rates should be specified.

12.3 Modification of OECD TG 203

Considerations:

- It is deemed important to consider issues on incipient LC_{50} determination for revising OECD TG 203 for hydrophobic substances.

Recommendations:

- Additional guidance is needed on range-finding – perhaps using the FET.
- The Rufli & Springer reduced test design approach (submitted to Environ. Toxicol. Chem.) should be considered.
- Replicates should be clearly defined as the test vessel because of binomial over-dispersion.
- Better and more consistent descriptions are needed. All sublethal effects should be reported.
  - The term *moribund* should be more clearly defined. The use of moribund instead of death in LC_{50} calculation should be considered. H. Rufli performed an analysis of how this would affect the LC_{50} values (see chapter 5; this study has not yet been published, but would need to be if considered in any revision).
- The 0 % and 100 % mortality need to be more clearly explained in the guideline (as non-mandatory).
  - There is a need to specify that enough data points are required such that a good estimate of the LC_{50} and slope of the dose-response curve can be made.
  - No additional test concentrations are needed, if the LC_{50} value can be calculated using five concentrations (e.g., there is no need to test an additional concentration just to make sure of obtaining 100 % or 0 % mortality).
  - If 0 % and 100 % mortality are reached with five concentrations, this should be reported, but should not be listed as a mandatory requirement in the TG.

13 When is it appropriate for a fish lifecycle test to be triggered from a fish short term screening assay (OEC TG 229), if there are no endocrine effects, but if some form of "reproductive effect" is observed?

Considerations:

- A positive result in OECD TG 229 is likely to require additional testing of some form (e.g., partial or full life-cycle testing) or reliance on existing data.

- The decision that directs the choice of additional test(s) if needed will be based on a weight-of-evidence approach and will depend on nature of response, exposure/bioaccumulation considerations, existing fish data and information from other taxa if available.
- Currently, there is no OECD TG or internationally standardized method available for conducting an intermediate level reproduction fish test, possibly in lieu of a fish full life-cycle test.

Recommendations:

- There is scope for developing a partial life-cycle test that includes the reproductive phase. This would be an enhanced OECD TG 229, with more concentrations, additional replicates and longer duration as a possible alternative test to a fish full life-cycle test (FFLC) or Japanese medaka multi-generation test (MMT).
- The review/analysis of existing fish data to inform the further development of a partial life-cycle reproduction test based on OECD TG 229 in a more comprehensive definitive design is recommended.
- In case this recommendation results in an actual OECD project, it will be important during the validation to consider the comparative sensitivity of fecundity measures between an enhanced OECD TG 229 and a FFLC.

14 Can (and when) could the following methodologies be employed in a fish testing strategy?

Considerations:

- Use of the limit test.
- Use of the threshold approach.
- Use of sequential (step-down) approach vs. effective range-finding.
- Use of screening methodologies that do not utilize animals (such as (Q)SAR tools, *in vitro* assays, or read-across.
- In general terms, there is a need to further optimise and reduce the number of test animals.

Recommendation:

- Adoption of limit testing has already been incorporated in OECD TGs, and it is recommended that further exploration of methods to reduce the number of fish used in existing (e.g., OECD TG 210) and future TGs (e.g., sequential testing and range finding) should be made.

15 Would an evaluation / discussion of the pros and cons of the various alternative methods be helpful?

Considerations:

- There is a need for more rapid assessment of such methods and a further reduction of the number of animals used.
- Potential problems with CBI may be solved similarly to how EcoSAR was developed for fish acute toxicity predictions. The International Life Sciences Institute/Health & Environmental Sciences Institute (ILSI/HESI) held a workshop on this subject in June 2010 and should be consulted.

Recommendation:

- There is a need to collate high quality fish chronic data sets (e.g. life-cycle test data, early life-stage test data) to help in the development of (Q)SAR and other computational and *in vitro* methods for use in the investigation of toxicity pathways relevant to chronic fish toxicity.

16 How can other data (e.g. from mammals, invertebrates, *in vitro* etc.) be utilized to support testing?

Considerations on possible sources of data:

- There is a need for information about modes of action (MOA) that could help target testing (various technologies/methods would apply).
- Weight-of-evidence analysis can be used to help inform concentration selection.
  - (Q)SAR models, log P information, BCF data, etc. should be utilized to help extrapolate from mammalian to fish data (i.e., from mg/kg to mg/L).
- Invertebrate data could be used for estimation of bioconcentration and acute toxicity.
- Mammalian systems might help to inform about the metabolism of certain chemicals – this is especially relevant to the strength of suspicion about an endocrine disruption mode of action, which is one of the key decision points in the testing strategy framework diagram.
- *In vitro* tests (cell lines, etc.) might help concentration-setting for acute tests.
- Use of "omics" might help to identify modes of action and/or additional endpoints (e.g., metabolomics could help to predict apical endpoints).
- OECD's approach for (Q)SARs (*via* the toolbox), as well as that of the toxicogenomics group, are good approaches that could be applied in other areas.
- Drawing upon internal screening / prioritization methodologies utilized by companies internally would be an effective way to identify potential alternative strategies.
- For replacement tests, there must be full validation.
- Other types of alternatives might have utility even without complete validation (reduction, refinement of the "3Rs").
- Animal use should be reduced by better targeting of species selection and test design.

Recommendations:

- A review document on the practical applications of mode of action and pathways that can be used to avoid unnecessary testing should be developed.
- For example, guidance could be elaborated on compounds or functional groups that target protein receptors (e.g. estrogen receptors) or compounds that interfere with metabolism – which is relevant to interspecies differences in fish.

17 Several guidelines allow use of alternative species – regional preferences notwithstanding, can technical guidance be developed/provided concerning the "best" (most appropriate) species to use for a given endpoint/risk concern?

Conclusion:

- Guidance can be developed for identifying the appropriateness of various fish species for OECD TGs.
- Selection of the most appropriate species will be important for optimizing the application of the TGs and will reduce the need for potentially redundant testing.

Recommendations:

- A review should be performed to consider the appropriateness of the various recommended/optional fish species in existing TGs and TGs in development considering endpoints and regulatory needs.
- This review should consider:
  - ease of testing, practicality;
  - ability to obtain and rear test animals;
  - sensitivity;
  - basic knowledge necessary on biology including growth rate, genetic sex markers etc.)
  - duration of life-stages;
  - capacity for biotransformation;
  - endpoint relevance, measurement of endpoints;
  - ecological and geographical relevance;
  - conservation status.

18 What are the options for reducing animal use in fish toxicity tests? Are there further options?

Considerations:

- The number of fish per tested concentration/control should be minimized by adjusting the required precision of the test result to biological/regulatory needs.

- The number of test concentrations and the number of animals per concentration required by the test guideline should be minimized.
  - *Triggers that would allow reduction of the total number of animals used:*
    - *range-finding;*
    - *slope of the dose-response curve for invertebrate tests; however, this might not be possible for certain modes of action.*
- Species specificity with respect to the number of animals should be considered (e.g, hatching success).
- Optimized range-finding (e.g., use of FET) and use of supporting information (e.g., from invertebrate tests) might further help in reducing the number of vertebrates used.

19 When should additional information be collected from *in vivo* tests?

Consideration:

- This information might be useful for the weight-of-evidence provisions in REACH Annex XI (General Rules for Adaptation of the Standard Testing Regime).

20 Can (and if so, when can) the following methodologies be employed in a fish testing strategy?

General considerations/conclusions:

- In order to be more widely applicable, retrospective analysis of data is required (specifically, for the threshold and step-down approaches).
- When moving through a testing framework, it should be considered whether certain methodologies are sufficient for a given circumstance before imposing additional test requirements.
- Increased consideration of exposure and fate information and prediction is needed.
- Limit test:
  - There is a need to clarify how existing data on invertebrates or existing chronic fish tests might be utilized to determine whether a limit test might be applied.
- Threshold approach:
  - Current guidance and use in REACH is only for acute use – there is no such test for chronic use.
  - Application of the threshold approach would be difficult for pesticides in the US, because levels of concern are different for different species (e.g., algae LOECs are different to those for fish, and the dose-response would be expected to be different).
- Step-down approach:
  - It should be noted that the step-down approach was not supported by the WNT.

- Screening methodologies that do not utilize animals (such as (Q)SAR tools, *in vitro* assays, or read-across) should be used.
  - It needs to be considered whether these methods are sufficient for a given circumstance before imposing (additional) test requirements.
- Fish embryo test (FET):
  - The fish embryo test (FET), which is currently undergoing validation, is another alternative methodology that should be considered in a testing framework in the future.