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

18-Aug-2005

English - Or. English

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

OECD SERIES ON TESTING AND ASSESSMENT
Number 34

GUIDANCE DOCUMENT ON THE VALIDATION AND INTERNATIONAL ACCEPTANCE OF NEW OR UPDATED TEST METHODS FOR HAZARD ASSESSMENT

Patric AMCOFF
Tel: +33 (0)1 45 24 16 19; Fax: +33 (0)1 44 30 61 80; Email: patric.amcoff@oecd.org

JT00188291

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

ENV/JM/MONO(2005)14
Unclassified

English - Or. English

OECD Environment, Health and Safety Publications
Series on Testing and Assessment
No. 34

GUIDANCE DOCUMENT ON THE VALIDATION AND INTERNATIONAL ACCEPTANCE OF NEW OR UPDATED TEST METHODS FOR HAZARD ASSESSMENT

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris

18 August 2005

Also published in the Series on Testing and Assessment:

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

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

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

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

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

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

No. 7, G*uidance 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, *Draft Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (1999)*

No. 20, *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, *Draft 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 Cell Transformation Assays for Detection of Chemical Carcinogens (draft)*

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 Validation and International Acceptance of New or Updated Test Methods for 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 Disruptors Testing and Assessment (EDTA) to Standardise and Validate the Utertrophic Assay (2003)*

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

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 Reproductive Toxicity Testing and Assessment (draft)*

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

No. 45, *Guidance Document on the Use of Mulimedia Models for Estimating Overall Environmental Persistance 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. 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 from the OECD/ICPS Workshop on Toxicogenomics, Kyoto, October 2004 (2005)*

© OECD 2005 ***Applications for permission to reproduce or translate all or part of this material should be made to: Head of Publications Service, OECD, 2 rue André-Pascal, 75775 Paris, Cedex 16, France.***

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 30 industrialised countries in North America, Europe and the Pacific, 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 subsidiary groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and subsidiary groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in nine 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; and Emission Scenario Documents.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/ehs/).

This publication was produced within the framework of the Inter-Organisation Programme for the Sound Management of Chemicals (IOMC).

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, OECD, UNEP, UNIDO, UNITAR and WHO. The World Bank and UNDP are observers. 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) 45 24 16 75

E-mail: ehscont@oecd.org

PREAMBLE

1. The development of this Guidance Document started in 1998 as a follow-up to the 1996 Solna Workshop on "Harmonisation of Validation and Acceptance Criteria for Alternative Toxicological Test Methods". Whereas the principles and criteria for validation and regulatory acceptance of new and revised test methods, agreed in Solna and included in the report of the Workshop (1), were widely accepted, it appeared that there was still a need to expand on these principles and provide more guidance on validation in general. Moreover, practical experience with validation studies has substantially accumulated since the Solna Workshop, and using this experience to provide more specific guidance was considered very useful.

2. A crucial step in the initial development of this Guidance Document was the OECD Secretariat's consultation with a selected number of internationally recognised experts in the area of test validation. The contributions from these experts, both in terms of practical advice, suggestions and final development of the first draft Guidance Document on "The Development, Validation and Regulatory Acceptance of New and Updated Test Methods in Hazard Assessment" have been extremely helpful. The draft Guidance Document was circulated to Member countries and other stakeholders in late 2001 and the OECD Secretariat received comments from eight member countries in addition to comments from EC, ICAPO and BIAC.

3. An OECD Conference on "Validation and Regulatory Acceptance of New and Updated Methods in Hazard Assessment" was held in Stockholm, Sweden in March 2002. The stated purpose of the conference was to develop, and achieve consensus on, practical guidance on principles and processes for the validation and acceptance of animal and non-animal test methods for regulatory hazard assessment purposes. This consensus guidance would be used to revise and further improve the draft OECD Guidance Document. The Stockholm Conference agreed upon 40 specific recommendations on how to improve and extend the draft Guidance Document (2). In addition the Conference agreed upon a number of general suggestions and editorial corrections for the upcoming revision taking into consideration the comments received after the first circulation round of the draft Guidance Document.

4. Two of the most controversial areas that many participants of the Conference felt were not properly covered were the issues of Data Interpretation Procedures (DIP) and the Use of Human Data. Therefore, the Conference made a strong recommendation that two separate OECD Workshops should be organised to further examine these issues. With the purpose of speeding up the finalisation of the draft Guidance Document, the Conference decided that a small Drafting Group should be established as soon as possible after the Meeting to initiate the revision. The group comprised representatives from the EU, Japan and North American Regulatory Communities as well as representatives from BIAC and ICAPO. In December 2002, the Secretariat established the Drafting Group through an official nomination round and in total eight experts offered their assistance. The Drafting Group had its first telephone conference in February 2003, during which the outline and approach of the revision was discussed and several topics that deserved extra attention were identified. These so called "hot-spot issues" were: (i) validation of (Q)SARs; (ii), Data Interpretation Procedures (DIP); (iii), the use of human and existing data; (iv), distinctions between different phases of the validation process; (v), retrospective validation and validation by convention; and, (vi) development of Test Guidelines from validated protocols.

5. The members of the Drafting Group came together in March 2003 in Paris to start the revision of the draft Guidance Document. The Drafting Group discussed and agreed on how to organise the work taking into account the outcome from the Stockholm Conference and the previous commenting round of the draft. The revised 2nd version was circulated to the WNT for comments in October 2003, and again, a considerable number of comments were received. Given the scope and extent of the comments, the Secretariat asked the WNT16 for guidance on how to continue the revision of the document. The US offered to host an Expert Consultation meeting in Bethesda, Maryland, to resolve these last topics. The

meeting was held on October 13-15, 2004, and successfully addressed the comments received in addition to making a number of rather extensive amendments to the draft Guidance Document.

6. One of the recommended follow-up activities of the Stockholm Conference was a workshop on DIPs, and after a kind offer from Germany (ZEBET) a workshop was held in Berlin on 1-2 July, 2004. The main purpose being to try to achieve consensus on principles for validation, and use of DIPs and Prediction Models (PMs) that are applicable to both *in vivo*, *in vitro* and ecotoxicological test methods. The workshop discussed and agreed upon a number of items, including terminology, and provided further guidance for the revision of draft Guidance Document 34. However, a consensus on the applicability of DIPs and PMs for different types of methods was not achieved. A Meeting Report is available (3).

7. Many experts have participated in the OECD meetings since the Solna Workshop in 1996 and the Secretariat would especially like to mention the experts that have been actively involved in the development and drafting of this document, including:

– Hans Ahr, Bayer AG, Wuppertal, Germany
– Michael Balls, EC/ECVAM, Ispra, Italy
– Robert Boethling, US-EPA, Washington, DC, USA
– Dorothy Canter, US-EPA, Washington, DC, USA
– Mark Chamberlain (BIAC) Unilever, UK
– Alan Goldberg, CAAT, Baltimore, MD, USA
– Petra Greiner, UBA, Germany
– Kailash Gupta, CPSC, Bethesda, MD, USA
– Karen Hamernik, US-EPA, Washington, DC, USA
– David Hattan, FDA, Bethesda, MD, USA
– Abigail Jacobs, FDA, Rockville, MD, USA
– Manfred Liebsch, ZEBET, Berlin, Germany
– Kimmo Louekari, Product Control Agency for Welfare and Health , Finland
– Yasuo Ohno, NIHS, Tokyo, Japan
– Willie Owens, (BIAC) Procter and Gamble, USA
– Richard Phillips, Exxon/Mobile, East Millstone, NJ, USA
– Amy Rispin, US-EPA, Washington, DC, USA
– Andrew Rowan, Humane Society of the US, Washington, DC, USA
– Len Schechtman, FDA, Rockville, MD, USA
– Jerry Smrchek, US-EPA, Washington, DC, USA
– Horst Spielmann, ZEBET, Berlin, Germany
– Martin Stephens (ICAPO), USA
– William Stokes, NIEHS, Research Triangle Park, NC, USA
– Gary Timm, US-EPA, Washington, DC, USA
– Leslie Touart, US-EPA, Washington, DC, USA
– Neil Wilcox, formerly of FDA, Rockville, MD, USA
– Marilyn Wind, CPSC, Bethesda, MD, USA
– Andrew Worth, JRC-EC, Italy
– Errol Zeiger, formerly of NIEHS, Research Triangle Park, NC, USA

The OECD Secretariat gratefully acknowledges these experts and others who have contributed to the project for their professional assistance and their indispensable contributions to the finalisation of this important Guidance Document.

8. This Guidance Document should be considered in conjunction with Guidance Document No.1 (Development of OECD Guidelines for Testing of Chemicals (4)) published in this same series of Guidance Documents on Testing and Assessment.

TABLE OF CONTENTS

PREAMBLE 9

I. INTRODUCTION 13

II. DEFINITION OF THE TEST METHOD 17

III. APPROACHES TO VALIDATION 20
Prospective Validation Studies 20
Retrospective Assessment of Validation Status 21
"Modular" Approaches to Assessing Validation Status 22
Performance Standards for Test Methods 24
Use of Performance Standards for Catch-up Validation 24
Use of Performance Standards for Modified Test Methods 25
Validation of Patented Methods 25
Validation of Test Batteries/Testing Strategies 25
Validation of (Quantitative) Structure-Activity Relationships - (Q)SARs 26

IV DESIGN AND CONDUCT OF VALIDATION STUDIES 27
General Considerations 27
Management of Validation Studies 31
Statistical Expertise 31
Conduct of the Prevalidation 32
Project Plan 34
Participating Laboratories 34
Reference Data 35
Selection of Test Chemicals and Chemical Management 36
Coding and Distribution of Test Samples 38
Test Method Decision Criteria and Data Interpretation 39
Monitoring of Participating Laboratory Performance 39
Inter-laboratory Testing 40
Data Collection 40
Data Analysis 41
Reporting 42
Record-Keeping/Data Dissemination 42

V. INDEPENDENT EVALUATION OF A VALIDATION STUDY (PEER REVIEW) 43
Mechanisms for Peer Review 43
Selection of Peer Reviewers 44
Charge to Peer Reviewers 45
Peer Review Process 45

VI. INTERNATIONAL REGULATORY ACCEPTANCE OF VALIDATED TESTS 47
Validation Study Outcomes 47
Criteria for Regulatory Acceptance 47
From Protocol to Test Guideline 51

VII. NEW TEST SUBMISSIONS: SUPPORTING DOCUMENTATION 50
Introduction and Rationale for the Proposed Test Method 50
Test Method Protocol Components 50

Characterisation and Selection of Substances Used for Validation of the Proposed Test Method.....51
In Vivo Reference Data Used to Access the Accuracy of the Proposed Test Method.....51
Test Method Data and Results.....53
Test Method Relevance (Accuracy).....53
Test Method Reliability (Repeatability/Reproducibility).....54
Test Method Data Quality.....54
Other Scientific Reports and Reviews.....54
Animal Welfare Considerations (Refinement, Reduction and Replacement).....54
Practical Considerations.....55
References.....55
Supporting Materials.....55

VIII. REFERENCES.....56

ANNEX I: Definitions and Glossary.....62

ANNEX II: Examples of validation studies within different areas of test method development.....69

Example 1: The OECD *Lemna* growth inhibition test. Development and ring-testing of Draft OECD test Guideline.....72
Example 2: Report of the Final Ring Test of the *Daphnia magna* Reproduction Test.....75
Example 3: Validation, Special Study and "catch-up" Validation of *In vitro* Tests for Skin Corrosion Potential – The Human Skin Model Test.....79
Example 4: The Local Lymph Node Assay (LLNA).....86
Example 5: The Up-and-Down Procedure for Acute Oral Toxicity (TG 425).....93

TABLES AND FIGURES

Table 1: Principles and criteria for test method validation. 23
Figure 1: From test development to validation: entry points for test method optimisation 28
Figure 2: Key factors for validation and regulatory acceptance of new and revised toxicology test methods 29
Figure 3: Role and interactions of the validation manager/management team 33
Table 2: Principles and criteria for regulatory acceptance of a new test method. 48
Table 3: Characterisation of chemicals tested. 51
Table 4: Test method accuracy assessment. 53

I. INTRODUCTION

9. Test Guidelines, including those produced by the OECD, are used by governments, industry and independent laboratories to determine the hazard or safety of chemicals. The use of Test Guidelines that are based on validated test methods promotes the generation of dependable data for human and animal health and environmental hazard assessment. It has been recognized that there is a need to develop guidance for the international community that outlines general principles, important considerations, illustrative examples, potential challenges and the results of experience gained in the area of test method validation. This guidance document has been drafted by representatives of OECD Member countries, based on advice from Member countries and OECD stakeholders.

10. The purpose of this document is to provide guidance on issues related to the validation of new or updated test methods consistent with current approaches. This document provides a synopsis of the current state of test method validation, in what is a rapidly changing and evolving area. Guidance on the more general aspects of OECD Test Guideline development is provided in the OECD Guidance Document for the development of OECD Test Guidelines for the testing of chemicals (4).While the principles of validation as described in this document were written for biology-based tests, they may be applicable to other Test Methods.

11. Test method validation is a process based on scientifically sound principles (5)(6) by which the reliability and relevance of a particular test, approach, method, or process are established for a specific purpose. Reliability is defined as the extent of reproducibility of results from a test within and among laboratories over time, when performed using the same standardised protocol. The relevance of a test method describes the relationship between the test and the effect in the target species and whether the test method is meaningful and useful for a defined purpose, with the limitations identified. In brief, it is the extent to which the test method correctly measures or predicts the (biological) effect of interest, as appropriate. Regulatory need, usefulness and limitations of the test method are aspects of its relevance. New and updated test methods need to be both reliable and relevant, *i.e.*, validated.

12. A set of principles for validation, also called the “Solna Principles”, were developed at an OECD Workshop in Solna Sweden in 1996, where it was agreed that these Principles apply to the validation of new or updated test methods for hazard assessment, whether they are *in vivo* or *in vitro*, or tests for effects on human health or the environment (1). These Principles are:

1) A rationale for the test method should be available. This should include a clear statement of scientific need and regulatory purpose.

2) The relationship of the endpoint(s) determined by the test method to the *in vivo* biological effect and to the toxicity of interest should be addressed. The limitations of a method should be described, *e.g.*, metabolic capability.

3) A formal detailed protocol must be provided and should be readily available in the public domain. It should be sufficiently detailed to enable the user to adhere to it, and it should include data analysis and decision criteria. Test methods and results should be available preferably in an independent peer reviewed publication. In addition, the result of the test should have been subjected to independent scientific review.

4) Intra-test variability, repeatability and reproducibility of the test method within and amongst laboratories should have been demonstrated. Data should be provided describing the level of inter- and intra-laboratory variability and how these vary with time.

5) The test method’s performance must have been demonstrated using a series of reference chemicals preferably coded to exclude bias.

6) The performance of test methods should have been evaluated in relation to existing relevant toxicity data as well as information from the relevant target species.

7) All data supporting the assessment of the validity of the test methods including the full data set collected in the validation study must be available for review.

8) Normally, these data should have been obtained in accordance with the OECD Principles of Good Laboratory Practice (GLP).

13. Given the continuing increase in the numbers and types of test methods being developed for varying purposes, the validation process should be flexible and adaptable. The extent to which these validation principles are addressed will vary with the purpose, nature, and proposed use of the test method. There are differences between *in vivo* assays and *in vitro* or *ex vivo* assays which should be considered in applying the validation principles. (7) In general, the closer the linkage between the effect measured and the toxicological effect of interest, the easier it will be to establish the relevance of the assay. The more closely a test measures/observes (an) effect(s) identical to the effect(s) of concern, the greater the confidence that the test will accurately characterize or model the effect in the target species of concern. (8)

14. There have been significant advances in test method development for toxicology and ecotoxicology over time. Many toxicity and ecotoxicity test methods have been in routine use for many years and have yielded, and continue to yield, information relevant to the toxicity of chemicals. Regulatory authorities and the regulated community accept these test methods for hazard and risk assessment purposes without the prerequisite for formal validation of those test methods using present-day validation criteria. OECD Test Guidelines are largely based upon such test methods.

15. Many alternative tests for potential human health hazards have been subjected to collaborative validation studies that preceded the systematic harmonization of validation principles by national or supra-national organisations/agencies (*i.e.*, ICCVAM, the Interagency Coordinating Committee on the Validation of Alternative Methods and ECVAM, the European Centre for the Validation of Alternative Methods). These test methods are accepted by international and national testing organizations and regulatory authorities for hazard and risk assessment based upon their history of use and proven utility. They provide a strong scientifically-based foundation for regulatory decision-making.

16. A number of these human health effect test methods are now used as the standards against which newly developed alternative test methods are evaluated. The ecotoxicological community has also conducted validation studies (often referred to as “ring-tests”) for decades to determine the reliability of test methods. Their focus has been on standardization of protocols and demonstrations of their transferability and reproducibility.

17. New test methods undergo validation to assure that they employ sound science and meet regulatory needs. Such new test methods have been developed for reasons that include human and ecological health concern, animal welfare considerations, data quality assurance, and to incorporate new and improved technologies. Increasing concern about the use of laboratory animals for toxicity studies and other effects of substances has led to widespread support of, and adherence to, the principle of the 3Rs of animal use in alternative test method development [Replacement, Reduction and Refinement (9)]. Regulatory authorities have endorsed the principle of the 3Rs. As a consequence, alternative test methods have been developed to replace the use of animals with non-animal systems, reduce the number of animals in a test, or refine the procedures to make them less painful or stressful to the animals under study (10).

18. With the introduction of new test methods, emphasis on continued test development and the need to work towards international acceptance of new or updated test methods, it has become apparent that a

means to promote international awareness of general principles, issues, and examples related to the validation of new test methods is needed.

19. This document provides guidance on the conduct of validation studies and how to demonstrate that the test methods have been appropriately validated. Validation studies can range from small-scale activities to large-scale multi-national programmes. Studies to support or effect validation can be performed under the auspices of the OECD, by individual Member countries, by test method sponsors, or by a number of different organisations. For example, some organisations are specialized in running test method validation of alternative test methods and/or are established by law as validation institutions, such as ECVAM at the European Union level, ICCVAM in the US and ZEBET (The German Centre for Documentation and Validation of Alternatives to Animal Experiments) in Germany1.

20. After a test has been validated and its performance (relevance and reliability) has been determined to be acceptable for its proposed use, a recommendation may be made that it be adopted as an OECD Test Guideline. Sponsors of test methods proposed for adoption as OECD Test Guidelines should provide evidence that the test method has been adequately validated in accordance with internationally recognized principles and criteria, such as those described by the Solna Workshop (1). These criteria apply for both new and revised tests. Regulatory acceptance is dependent upon the outcome of the validation, with consideration of the principles given above. In considering the regulatory acceptance of a test method, the following criteria are also important (1).

1) Application of the method provides data that adequately predicts the end-point of interest in that it demonstrates either a linkage between (i) the new test and an existing test method or (ii) the new test and effects in the target species.

2) The method generates data for risk assessment purposes that are at least as useful as, and preferably better than, those obtained using existing methods. This will give a comparable or better level of protection for human health or the environment.

3) There are adequate testing data for chemicals and products representative of the type of chemicals administered by the regulatory programme or agency (*e.g.*, pesticides, cosmetics).

4) The test should be robust and transferable and allow for standardisation. If highly specialised equipment, materials or expertise are required, efforts should be sought to facilitate transferability. This is an important criterion to be considered at an early stage of a validation study. [Note added by the Secretariat: According to current OECD policy, the test should not require equipment or material from a unique source. This would prevent the acceptance of patented methods. The Solna Workshop did not discuss the issue of patented tests but referred the issue to higher policy levels at OECD].

5) The test is cost effective and likely to be used.

6) Justification (scientific, ethical, economic) should be provided for the new method with respect to any existing methods available. In this respect due consideration should be given to animal welfare consideration including the 3Rs.

[1] There are other organisations which contribute greatly to the development and use of reduction, refinement and replacement alternatives by providing funding to research activities and/or organising activities that promote the principle of the 3R, such as the Johns Hopkins University Centre for Alternatives to Animal Testing (CAAT), the Fund for the Replacement of Animals in Medical Experiments (FRAME), the European Research Group for Alternatives to Animal Testing (ERGATT), the National Centre for Alternatives (NCA) in the Netherlands and the Swedish Animal Welfare Agency (SAWA).

21. Although the process of test validation, as described in this document, is separated into discrete phases the test method validation process is, in practice, a continuum between the various elements of validation. Each phase of the process uses and builds on results of other phases of the work. Regardless of whether the approach to validation is prospective, retrospective or modular (11), or otherwise, the goal of the validation process is to ensure that each of the OECD validation principles has been adequately addressed. Therefore, it is essential that documentation of adequate validation be provided as outlined and described in Chapter VII.

22. Definitions of specific words and terms used in this document are presented in Annex I, where examples of the variability of terminology used by different validation agencies are also briefly covered. For examples of different validation studies including *in vivo*, *in vitro* health and ecotoxicity studies, see annex 2.

23. This document addresses the important steps and aspects that should be considered prior to and during the validation process. They include:

(i) The definition of the test method and related issues (*e.g.*, purpose, decision criteria, endpoints, limitations);

(ii) the design and conduct of prevalidation studies leading to the optimisation of the test method;

(iii) the design and conduct of the formal inter-laboratory validation work, based on the outcomes of the prevalidation studies and aiming at accumulation of data on the relevance and reliability of the test method, and;

(iv) the overall data evaluation and subsequent validation study conclusion, keeping in mind the requirements of regulatory authorities, for submission of information relating to new or modified test procedures.

It also discusses the need for and the extent of an independent evaluation, or peer review, of test methods being validated. This document provides a synopsis of the current state of test method validation.

II. DEFINITION OF THE TEST METHOD

24. In the context of this Guidance Document, and of the overall OECD Test Guidelines Programme, a test method is an experimental system that can be used to obtain a range of information from chemical properties through the adverse effects of a substance. The term "test method" may be used interchangeably with "assay" for ecotoxicity as well as for human health studies.

25. The planning and conduct of a validation study encompass and are applied to a wide range of test methods and their procedures, which may be applied either as a stand alone test or as components of a test battery or tiered test system. These considerations may include chemical properties, (Q)SAR procedures and models, *in vitro* tests, *in vivo* animal tests, ecotoxicity and ecological test methods, and available human data and test methods. In the future, new technologies such as genomics, proteomics and other novel techniques may be incorporated into test methods. These test methods also vary in having single and multiple endpoints and having different types of endpoints (*e.g.*, cytotoxicity, tissue and organ weights, histopathology grades, and serum and clinical analyses). Validation studies may address new tests, substitute tests, modifications of endpoints in existing tests, and enhancement of existing test methods, regardless of whether tests are for chemical properties, human health or environmental effects. The amount of information required in these validation studies, the number of chemicals tested, when and to what extent to use testing, and the number of laboratories participating may vary as long as such factors do not adversely affect the outcome of the validation process. For these reasons, the planning and conduct of a validation study for a test method are expected to be done on a case-by-case basis.

26. The first step in developing a new or revised test method is the definition of the test method. The definition includes the test method's chemical, biochemical or biological basis. This includes a rationale for the relevance of the results produced such as the endpoints to be measured and a rationale or decision criteria for how the results are to be interpreted and used, *e.g.*, by using a data interpretation procedure (DIP) or a prediction model (PM)(3). The definition should also include the use of the results in the assessment of human health or environmental effects, and, as appropriate, the test method's position relative to other data/information requirements and test methods, *e.g.*, location in a testing battery or tiered testing strategy.

27. To elaborate on the above points, an example of the basis for a test can be described by the mechanistic or type of effects it is designed to measure (*e.g.*, inhibition of the differentiation of embryonic stem cell lines, induction of mutation by base-pair substitution) (12). Examples of the role or use of the test may be described in terms of:

- The biological effect it is designed to predict (*e.g.*, hypersensitivity potential in people based on a mechanism-based biological response in the murine Local Lymph Node Assay (13)).
- The biological effect it is designed to measure (*e.g.*, carcinogenicity in animals, toxicity in a specific species of fish, alterations in one or more component of the endocrine system to an extent that causes adverse health effects in the rat).
- The species of selection in biological tests is designed to represent e.g., a special ecological function found in healthy environmental communities, a special taxonomic group, trophic level or microhabitat.

Examples of a test method's scientific and regulatory needs include identifying whether a chemical might have an effect on the liver, or an effect on the number of eggs produced by fish species. Examples of a test method location in a tiered testing strategy might be as either *in vitro* or *in vivo* assays that might precede a long term study of reproductive effects. A test method's applicability domain refers for example to chemical classes, mechanisms of action and to the range of responses for which the test can be used reliably. The applicability domain may also specify known limitations of the test method such as

restrictions on the classes of substances that can be accurately identified or measured by the test. Furthermore, the test method may be suitable for identifying active substances, but not inactive substances, or vice versa.

28. The following are more detailed descriptions of five categories of test methods and testing strategies. In a strict sense, only the first three are categories of test methods and they may all be replacement test methods and/or part of a test battery.

- A screening test method is a rapid, usually simple test performed for the purposes of prioritizing or grouping substances in general categories of potential modes of action (*e.g., in vitro* binding to the oestrogen receptor). The results from screening tests are generally used for preliminary decision making and to set priorities for additional and more complex tests. Although the results from screening tests, alone, may not be sufficient for risk assessment purposes, there may be circumstances where such results may be combined with other test results in a tiered testing approach to provide in the hazard/risk assessments. As for all test methods, the limitations of the test also need to be taken into account, for example, certain screening tests may only be able to identify active substances, but not inactive substances. Other screening assays cannot distinguish agonist from antagonist activities (*e.g., in vitro* oestrogen receptor binding assays).
- A definitive test method generates sufficient data to characterise the specific hazard of the substance without the option to always require further testing (*e.g.,* two-generation reproductive toxicity study). Ideally, it is a test method which provides sufficient stand-alone information on dose response and adverse effects to permit risk assessment and risk management organisations to make decisions to reduce or minimize hazard and risk. However, in practice, data from other test methods and experiments, if available, are typically included in risk assessment and management, and the definitive test results do not stand alone.
- An adjunct test method provides data that add to the data set or help interpret the results of other test methods to assist the assessment process (*e.g.,* toxicokinetics of an animal carcinogen to interpret the carcinogenicity findings in different species).
- A replacement test method is designed to replace an existing test method, whether it is a screening or definite test. For a test method to be considered as a replacement, it should offer advantages over the accepted test method. Such advantages may be a reduction in animal numbers or suffering (*i.e.,* consideration of the principle of the 3Rs), equal or more accurate prediction of human health and environmental hazards, increased sensitivity or reliability over an existing test method, increased information useful for support of decisions, decreased cost, or enhanced ease of use. The validation exercises should be sufficiently rigorous to demonstrate unambiguously that the new or revised test procedure meets or exceeds the performance of the test method to be replaced.
- A test battery comprises a number of test methods that are generally performed at the same time or in close sequence (*e.g.,* genotoxicity test battery or aquatic base set tests that include data for algae, daphnia, and fish species). Each test method within the battery is designed to complement the other test methods such as measure a different endpoint or mechanism of genotoxicity or a different component of ecological systems. Component test methods of test batteries are treated as individual test methods for validation purposes and it is necessary to demonstrate that the combination of test methods produces reliable and relevant results and is more effective than the individual tests. In general, substitution of any component of the battery should improve its performance.

29. Current test methods may be periodically revised for reasons of enhancement and improvement to produce an enhanced test method. Examples of enhancement could involve the addition of new endpoints in order to increase the sensitivity of the test method or introduce augmentations of existing endpoints. New methods or endpoints could also be designed, for example, to reduce the number of animals or to allow the assay to be performed in less time. As for a new test method, validation of new endpoints that address new mechanisms and toxicities is needed so that the results can be interpreted with scientific and regulatory confidence. With existing complex test methods, the addition of new measurements or endpoints that increase the complexity, both in logistics and interpretation, require appropriate consideration during validation.

III. APPROACHES TO VALIDATION

30. To establish the scientific validity of a test method for a particular purpose, it is necessary to obtain information that fulfils the criteria and scientific principles for test method validation, and to assess the extent to which the principles have been fulfilled. The information can be obtained by consulting existing sources (retrospective assessment of validation status), by generating new experimental data (prospective validation studies), or by adopting a combination of prospective and retrospective validation approaches.

31. The principles for validation apply to all test methods. Scientific rigour is always required, regardless of the scope of the validation, the type of test method, or whether the method is new, revised or historical. Nevertheless, the level of necessary assurance that is appropriate for a specific purpose varies and needs to be identified on a case-by-case basis.

32. The amount and kind of information needed and the criteria applied to a new test method will depend on a number of factors. These include:

- The regulatory and scientific rationale for the use of the test method;
- the type of test method being evaluated (*e.g.*, existing test, new test);
- the proposed uses of the test method (*e.g.*, mechanistic adjunct, screening or definitive test, total or partial replacement of an animal or non-animal test method);
- the proposed applicability domain of the test method (*e.g.*, restricted chemical classes, multiple chemical classes);
- the relationship of the test species to the species of concern;
- the mechanistic basis of the test and its relationship to the effect(s) of concern; and,
- the history of use of the test method, if any, within the scientific and regulatory communities.

33. For cases in which a new test method addresses an endpoint that has not been previously considered or for which there are no or inadequate established reference data, a rigorous assessment of the predictive capacity may not be possible, however, the relevance should be assessed using all available information. Such tests that have shown to be reliable, but whose relevance has only been partially assessed by comparison with related reference data (*e.g.*, TG 414 (14) and TG 424 (15)), should preferably have their validation status periodically reviewed, with full consideration of all relevant available data. In addition, in cases where it is contemplated enhancing an assay by adding new endpoints onto an existing complicated study design some degree of validation is needed to demonstrate such additions are logistically feasible, would not impair existing procedures and measurements and are not redundant.

Prospective Validation Studies

34. A prospective approach to validation studies should be carried out when some or all of the information considered necessary to adequately assess the validity of a test method is not available, so that new experimental work needs to be performed. The data generated are used to determine the performance characteristics (*i.e.*, relevance, intra-, and inter-laboratory reproducibility), advantages, and limitations of the test method. Following completion of a prospective validation study, an evaluation of all the available information and data would then be conducted in order to assess the validity of the test method for a specific proposed use.

35. Before resources are invested into a prospective validation study (16) certain types of information should ideally be available, including:

- A clear statement of the scientific rationale and regulatory purpose of the test method.
- An explanation of the need for the test method. The most obvious reason for the development of a new test method is to address an area of toxicological concern for which tests do not yet exist. Another reason is that the new test method or a new test species will have an advantage(s) over existing test methods, such as its speed or ease of performance, reduced cost, reduction in numbers of animals utilised, refinement (*i.e.,* reduced stress and pain in animals), increased sensitivity and accuracy, improved identification of specific classes of substances, or its use as an *in vitro* screening test or replacement for an *in vivo* test.
- Clear and comprehensive standardized test method protocols, preferably in compliance with GLP Principles, together with standard operating procedures (SOP), as appropriate. This should include a description of the test system, exposure conditions, dose selection procedures, endpoint(s) assessed, measurements taken, specialized equipment or supplies that may be needed, measures of variability, the way in which the results are calculated and expressed, and the use of positive and negative controls and other performance checks.
- A description of the rationale and criteria to be used to analyse and interpret the test results. (*e.g.*, decision criteria, PM or DIP, where appropriate).
- Background information supporting the expected or established usefulness and limitations of the selected test method. Included with this information should be a list of available publications and unpublished reports on the test method.

Retrospective Assessment of Validation Status

36. A retrospective assessment of the validation status of a test method can be carried out by reviewing all of the information and data available that supports or questions the validity of a test method, including the results of any previous validation studies that have been conducted. In such a case, new experimental work may not need to be performed, but additional data analysis and evaluation work may be necessary. Whether or not additional validation is conducted should be decided on a case-by-case basis, and consistent with the principles outlined in this document.

37. For test methods in which there are published data available, the first effort should be to conduct a thorough examination of the peer-reviewed scientific literature, and any other relevant and credible reports and publications for information about the performance of the test method. This effort should be undertaken to determine whether sufficient information is available in the literature to substitute fully, or for some components, in a prospective validation process. In some cases the published information is limited in description or in the presentation of test validation information, due to journal space limitations and costs. Also, the validation data may be only available in the "grey literature", *e.g.*, in technical reports, draft papers, and other miscellaneous unpublished documents that are difficult to obtain. Furthermore, these documents may be written in a language which would require translation. In all cases the principles as articulated in Table 1 should be fulfilled for test methods without a full prospective validation study, but instead, with a history of use and published data in the scientific literature. These principles are needed to determine the performance characteristics, advantages, and limitations of any tests proposed for regulatory adoption. Retrospective assessment of the validation status of these test methods will require obtaining the study protocols and independently analyzing raw data so that these critical features of validation can be adequately evaluated. For such a case, it may be useful to form an independent expert panel to evaluate the assembled data according to the validation principles described above, and to assess the reliability and relevance of the test, and its advantages and limitations, without undertaking additional laboratory testing. There may be situations, however, when data gaps need to be filled as part of the process to achieve adequate validation for a specific proposed use. In these cases, attempts should be made to limit the extent

of new animal testing as far as possible. Retrospective validation assessments should also be subjected to independent peer review (see chapter V; Independent evaluation of a validation study: peer review).

38. When conducting a retrospective assessment of the validation status of a test method, it may be necessary to address some special issues. For example, it may be necessary to determine the extent that experimental data obtained with different variants of a test method protocol can be pooled for the purposes of a single analysis, or whether an assessment of test method validity can be made in the absence of raw data.

"Modular" Approaches to Assessing Validation Status

39. The decision as to whether a retrospective assessment of validation status is adequate or whether prospective validation studies are necessary depends on whether or not adequate reliable information is available to substantiate the validity of a test method. In practise, it is often necessary to adopt a combination of retrospective and prospective approaches in order to generate adequate information and data.

40. A general conceptual framework for documenting the studies necessary to assess the validation status of a test method, called the "modular approach" to validation, has been proposed (11). In this approach, information needed to support the validity of a test method is organised into modules that provide the following information:

(i) Test definition (including purpose, need and scientific basis);

(ii) intra-laboratory repeatability and reproducibility;

(iii) inter-laboratory transferability;

(iv) inter-laboratory reproducibility;

(v) predictive capacity (accuracy);

(vi) applicability domain; and,

(vii) performance standards.

According to this proposal each of these "modules" of information should be available in order to assess the validation status of the test method for a specific proposed purpose, but they need not be completed in a linear/sequential fashion.

TABLE 1. PRINCIPLES AND CRITERIA FOR TEST METHOD VALIDATION

a) **The rationale for the test method should be available.**

This should include a clear statement of the scientific basis, regulatory purpose and need for the test.

b) **The relationship between the test method's endpoint(s) and the (biological) phenomenon of interest should be described.**

This should include a reference to scientific relevance of the effect(s) measured by the test method in terms of their mechanistic (biological) or empirical (correlative) relationship to the specific type of effect/toxicity of interest. Although the relationship may be mechanistic or correlative, test methods with biological relevance to the effect/toxicity being evaluated are preferred.

c) **A detailed protocol for the test method should be available.**

The protocol should be sufficiently detailed and should include, *e.g.*, a description of the materials needed, such as specific cell types or construct or animal species that could be used for the test (if applicable), a description of what is measured and how it is measured, a description of how data will be analysed, decision criteria for evaluation of data and what are the criteria for acceptable test performance.

d) **The intra-, and inter-laboratory reproducibility of the test method should be demonstrated.**

Data should be available revealing the level of reproducibility and variability within and among laboratories over time. The degree to which biological variability affects the test method reproducibility should be addressed.

e) **Demonstration of the test method's performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used.**

A sufficient number of the reference chemicals should have been tested under code to exclude bias (see paragraphs on “Coding and Distribution of Test Samples”).

f) **The performance of the test method should have been evaluated in relation to relevant information from the species of concern, and existing relevant toxicity testing data.**

In the case of a substitute test method adequate data should be available to permit a reliable analysis of the performance and comparability of the proposed substitute test method with that of the test it is designed to replace.

g) **Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of GLP.**

Aspects of data collection not performed according to GLP should be clearly identified and their potential impact on the validation status of the test method should be indicated.

h) **All data supporting the assessment of the validity of the test method should be available for expert review.**

The detailed test method protocol should be readily available and in the public domain. The data supporting the validity of the test method should be organised and easily accessible to allow for independent review(s), as appropriate. The test method description should be sufficiently detailed to permit an independent laboratory to follow the procedures and generate equivalent data. Benchmarks should be available by which an independent laboratory can itself assess its proper adherence to the protocol.

Performance Standards for Test Methods

41. The purpose of performance standards is to communicate the basis by which new test methods, both proprietary (*i.e.*, copyrighted, trademarked, registered) and non-proprietary can be determined to have sufficient accuracy and reliability for specific testing purposes. These performance standards, based on validated and accepted test methods, can be used to evaluate the accuracy and reliability of other analogous test methods (colloquially referred to as "me-too" tests) that are based on similar scientific principles and measure or predict the same biological or toxic effect (17). However, documentation as outlined in Chapter VII should be developed to adequately address to what extent the validation and acceptance criteria have been met. Performance standards such as those described below should be provided, as appropriate, in Test Guidelines issued for new test methods.

The three elements of performance standards are:

Essential test method components: These consist of essential structural, functional, and procedural elements of a validated test method that should be included in the protocol of a proposed, mechanistically and functionally similar test method. These components include unique characteristics of the test method, critical procedural details, and quality control measures. Adherence to essential test method components will help to assure that a proposed test method is based on the same concepts as the corresponding validated test method.

Minimum list of reference chemicals: These are used to assess the accuracy and reliability of a proposed, mechanistically and functionally similar test method. These chemicals are a representative subset of those used to demonstrate the reliability and the accuracy of the validated test method. To the extent possible, these reference chemicals should:

- Be representative of the range of responses and effects that the validated test method is capable of measuring or predicting

- Have produced consistent results in the validated test method and in the reference test method and/or the species of interest

- Reflect the accuracy of the validated test method

- Have well-defined chemical structures and purity

- Be readily available (*i.e.*, from commercial sources)

- Not be associated with excessive hazard or prohibitive disposal costs

These reference chemicals are the minimum that should be used to evaluate the performance of a proposed, mechanistically and functionally similar test method. If any of the recommended chemicals are unavailable, other chemicals for which adequate reference data are available could be substituted. To the extent possible, the substituted chemical(s) should be of the same chemical class and activity as the original chemical(s). If desired, additional chemicals representing other chemical or product classes and for which adequate reference data are available can be used to more comprehensively evaluate the proposed test method. However, these additional chemicals should not include any that had been used to develop the proposed test method.

Accuracy and reliability values: These are the comparable performance requisites that should be achieved by the proposed test method when evaluated using the minimum list of reference chemicals.

Use of Performance Standards for Catch-up Validation

42. Subsequent to the determination that a test method is valid for a specific proposed use, establishment of performance standards based on this test method may be used to assess the validity of

other structurally and functionally similar test methods. The identification and use of standard reference substances to evaluate these similar methods (16) can facilitate so-called "catch-up validation" (18)(19). This expedited approach has been successfully applied to the validation of a human skin model for skin corrosivity (20). Regardless of the approach used to assess the validation status of a test method, all new or revised test methods should undergo a formal peer review. Each test method will need to be considered on its own merits and should meet the general principles for test method validation.

Use of Performance Standards for Modified Test Methods

43. As a new test method is used, the number of results obtained with the test expands. These results should ideally be periodically reviewed to determine if the usefulness of the test method has changed. It may be appropriate from time to time to review and reassess the validation status of established test methods for hazard/risk assessment. Such reviews could be "event-driven" (in response to new scientific findings individually suggesting a need for reassessment) or periodic (to determine whether the collected body of new information amassed since the previous review warrants a reassessment).It is also important to be aware of possible changes over time from the original properties or responsiveness of the test system, or modification of procedures that might affect the outcome of the test method. Monitoring of positive, negative and benchmark controls can aid in determining the occurrence or effects of changes.

44. It may be appropriate to evaluate proposed potential improvements to an approved test method. Prior to adoption of such changes, there should be an evaluation to determine the effect of the proposed changes on the test's performance and the extent to which such changes affect the information available for the other components of the validation process. Depending on the number and nature of the proposed changes, the generated data and supporting documentation for those changes, they should either be subjected to the same validation process as described for a new test, or, if appropriate, to a limited assessment of reliability and relevance using established performance standards. The extent of validation studies that would be appropriate should be determined on a case-by-case basis. Major changes should be subjected to a new independent peer review. This should take into account that the test was previously adopted, and that the specific changes and enhancements (or improvements) proposed are intended with the goal of improving the test's performance. In addition, consideration should be given to such factors as numbers of laboratory animals used, feasibility of the proposed revisions, cost and time effectiveness, and laboratory capacity.

Validation of Patented Methods

45. For validation purposes, patented or proprietary methods should be treated like other methods and they should be scientifically valid for their proposed specific use. The OECD currently will not develop Test Guidelines that require the use of a unique instrument or process owned by a patent. One reason for this is that the method should be readily available to all potential users: another is to avoid market monopoly of an OECD test method by a private company. One option to make it possible to adopt a patented test as an OECD Test Guideline would be to include a detailed generic description of the method and provide proper reference to the validated, patented version of the method, together with a set of performance standards (17). Any other version of the method, whether patented or not, should meet these performance standards.

Validation of Test Batteries / Testing Strategies

46. A test battery, base set or tiered testing approach is a series of tests usually performed at the same time or in close sequence. Each test in the battery is selected so as to complement the other tests, *e.g.*, to identify a different component of a multi-factorial effect, or to confirm another test. A battery of tests may also be arranged in a hierarchical (tiered) approach to testing. In a tiered approach, the tests are used

sequentially; the tests selected in each succeeding level are determined by the results in the previous level of testing. In these cases, a decision is made at each tier on whether there is sufficient information to stop testing or whether additional testing is needed, at present this is a common approach for *e.g.*, ecotoxicity testing.

47. Comprehensive guidance on the validation of test batteries and testing strategies has not been developed. However, a few general considerations are given here.

48. Individual tests within a battery of tests (or testing strategy) should be validated using the validation principles described in this document, taking into consideration their restricted roles in the test battery/testing strategy. Justification for the acceptance of a test battery should be primarily on the basis of its overall performance for its intended purpose. When tests are used in a tiered approach, the overall results will normally depend on the strength of the individual tests in the tier unless certain tests in the tier are used in a confirmatory manner or have a "safety net" function.

49. The performance of a tiered testing strategy, *i.e.*, its predictive capacity and ability to replace or reduce or refine the use of animals, may be evaluated by simulating possible outcomes of the strategy using existing data, as exemplified by assessments of tiered testing strategies for skin corrosion (21)(22)(23) and for eye irritation (24)(25).

Validation of (Quantitative) Structure-Activity Relationships – (Q)SARs

50. Structure-activity relationships and quantitative structure-activity relationships, collectively referred to as (quantitative) structure-activity relationships, (Q)SARs, are theoretical models that relate chemical structure to physicochemical properties, environmental fate parameters, or biological activity. Such models may provide useful predictions of such endpoints, and therefore reduce the need to obtain experimental data for the purposes of hazard assessment.

51. The specific issues associated with validation of (Q)SARs is addressed in a separate Guidance Document on (Q)SAR Validation that is being developed by the OECD, taking into account recent publications in the (Q)SAR field (26)(27)(28)(29)(30)(31)(32).

IV. DESIGN AND CONDUCT OF VALIDATION STUDIES

General Considerations

52. The purpose of the validation process is to determine the performance characteristics, usefulness, and limitations of a test method that is under consideration for use in a regulatory context, and to determine the extent that results from the test can be used for hazard identification, and to support risk assessments or other health and safety decisions. Validation is distinct from test method development.

53. The principles and criteria for validation are applicable to all test methods (see paragraph 17). Scientific rigour is always required, regardless of the scope of the validation, the type of test, or whether the method is new or revised. However, the level of assurance that is appropriate for a specific purpose and type of test varies and should be assessed on a case-by-case basis. Some general guidance may be given, particularly for certain types of studies.

54. The key factors and workflow in the validation process are shown in Figures 1 and 2. Regulatory authorities may have specific requirements for the analysis and summarisation of results of validation studies prior to the submission of such studies.

55. In certain situations an existing test that has been used as a research tool or in product development will be proposed for use in a regulatory context. Although the test may have been widely used in the past, there may not have been a systematic accumulation of data using a standardized protocol to allow a formal evaluation of its reliability and relevance. The validation of such a test may proceed by two different pathways, depending on the available database;

(i) The review of all available data so as to perform a retrospective analysis of the reliability and relevance of the test; or, if sufficient and appropriate validation data are not available,

(ii) by undertaking additional prospective validation studies, in order to develop a data set of relevant chemicals, to be considered together with previously existing data.

56. In some cases a combination of both approaches might be appropriate in order to provide the relevant information necessary to assess the validity of a test.

57. Regulatory acceptance is dependent on the outcome of scientific validation,. The acceptance process will be greatly facilitated by the early involvement of regulatory authorities in the planning and design of the validation study so that it will address areas of interest and concern to the respective regulators.

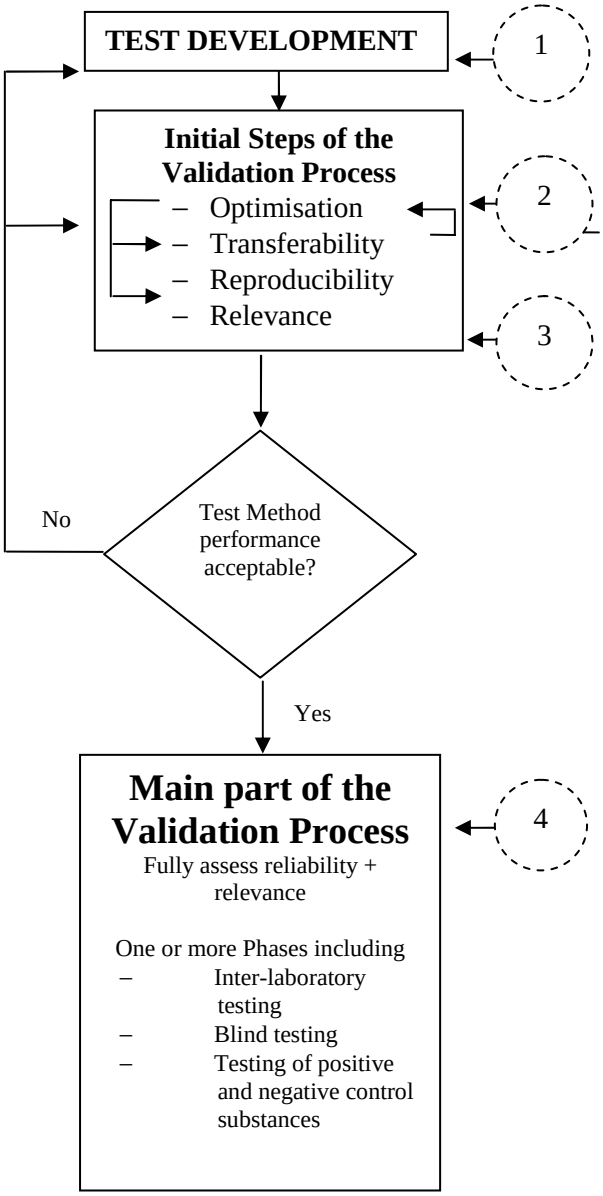
58. Before subjecting a test method to a formal inter-laboratory validation study, standardization and an initial assessment of the proposed test method is recommended. This initial phase is often referred to as “prevalidation” and is performed for several reasons (33). Three of the most important are:

(a) Refinement, optimisation and standardisation of the test protocol and standard operating procedures (SOPs), as appropriate, so that they can be readily used by other laboratories;

(b) development of preliminary data on reliability and relevance of the test method; and,

(c) avoiding the unnecessary expenditure of resources on multi-laboratory studies of methods of doubtful performance (Figure 1).

FIGURE 1. FROM TEST DEVELOPMENT TO VALIDATION: ENTRY POINTS FOR TEST METHOD OPTIMISATION



PREVALIDATION

Point 1

Each element of a test (for example chemical exposure time, histopathology, clinical chemistry, etc.,) needs to be carefully explored and evaluated to determine the optimum conditions/details.

Point 2

At least one laboratory independent from the test method developing laboratory will conduct the assay for an initial assessment and review of its inter-laboratory transferability and reproducibility. In case the test method fails to provide sufficient reproducibility, depending on the degree of failure, it may or may not be considered for further optimisation (33, 34).

Point 3

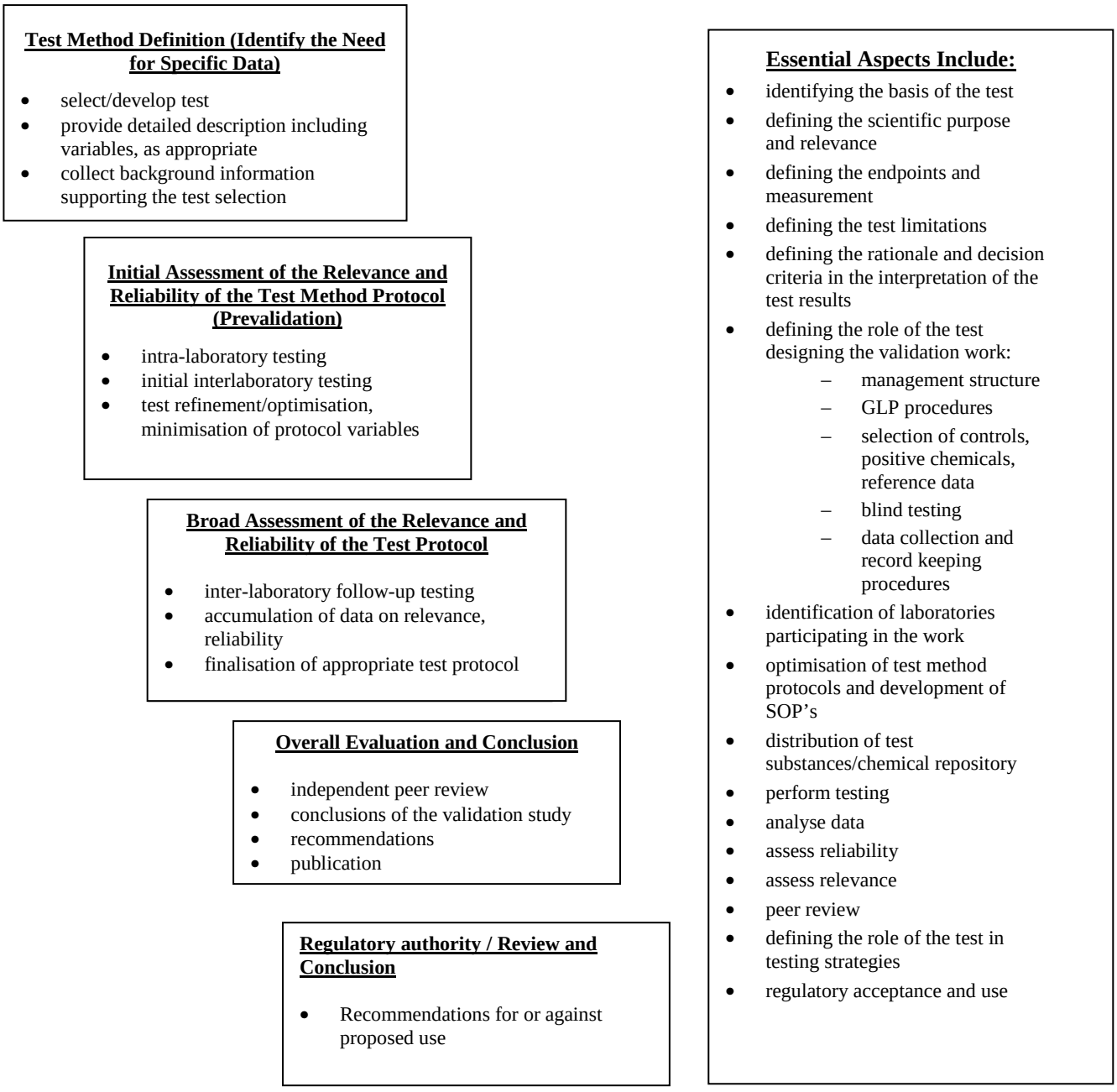
Test methods may fail to show acceptable performance in terms of relevance rather than a lack of reproducibility. This could be the result of a weakness of the test method protocol. If further test optimisation does not solve the performance issue, redesign of parts of the test method may be in order or a rejection of the proposed test method should be considered.

INTER-LABORATORY VALIDATION

Point 4

In certain cases where the result of the prevalidation was not sufficient to justify further validation, but was nevertheless promising, a test method could, after refinement, proceed to the main part of the validation study, provided this phase is properly designed to deal with the imperfections. In this case it is recommended to perform the main validation study in more than one phase.

FIGURE 2. KEY FACTORS FOR VALIDATION AND REGULATORY ACCEPTANCE OF NEW AND REVISED TOXICOLOGY TEST METHODS



59. Toward these ends the prevalidation process is also used to establish appropriate acceptable response ranges for positive and negative control substances to develop the data recording and evaluation criteria, to confirm or refine the preliminary decision criteria, to obtain a preliminary assessment of the transferability, reproducibility, and reliability of the test, and to identify limitations of the test. Attention should be paid to the latter point, because experience has shown that tests performing well in the developing laboratory, or when used with only a limited range of chemical substances, may not perform as well in a validation study as a result of limitations of the test that had not been determined and defined. Other reasons for performing a prevalidation could include an assessment of a number of protocol variations in order to select the most effective protocol(s) for the formal inter-laboratory validation study, or to clarify aspects of the test protocol that may not be clear from the available information.

60. After the initial evaluation has been completed, and prior to the planning and initiation of the formal inter-laboratory validation study, the results obtained should be fully evaluated, including an assessment of whether the conclusions of the study are supported by the data, and whether the proposed protocol is adequate and appropriate. Based on this review, a decision would be made to:

- Proceed to a formal inter-laboratory validation study using the selected test method protocol, SOPs (as appropriate), and data evaluation or decision criteria as appropriate to maximize reproducibility and accuracy.
- Recommend further research/development to improve/modify the test method.
- Proceed to an independent assessment of the test method, with the aim of having it accepted into regulatory practice in the absence of a formal inter-laboratory validation study. This option would be followed under circumstances where the results from the prevalidation study, pre-existing data, and, if applicable, understanding of the test mechanism could support a recommendation that no further validation work is necessary.
- Abandon the test method as not worthy of further validation.

61. After prevalidation is successfully completed the formal inter-laboratory validation process may be initiated (footnote; Ring-tests may be conducted in (a) prevalidation to develop the technical aspects of a test method or in (b) validation to evaluate the performance characteristics of a preferred method. The results of the prevalidation will be used in the design of the formal inter-laboratory validation study. A well planned and designed prevalidation could lead to reduced cost and effort during the latter multi-laboratory phase of the validation study.

62. Following conclusion of the validation process, there could be a National or regional regulatory acceptance process. For both sets of processes, there are principles and criteria that each test must meet (Figure 2). Table 1 lists the principles and criteria for a valid test method. These principles and criteria were adopted from a range of sources (1)(5)(6)(8)(16)(33)(35)(36). They provide a general description of the information needed at various points in the test development and validation process. At each stage of the process, from definition of the test method through the various stages of validation to preparation of the final report, the information provided should be sufficient to support a decision to proceed to the next stage.

63. One conceptual approach proposed for managing and tracking the data from validation of a test method is to collect the data and information in seven distinct “modules” as previously described in **paragraph 43** (11). Irrespective of the approach used, a test method submitted for review of its validation status should be accompanied by documentation outlined in Chapter VII (New Test Submissions: Supporting Documentation).

Management of Validation Studies

64. Validation studies are normally conducted under the auspices of a sponsor, such as international bodies, government entities and validation organisations for alternative methods (*e.g.*, ECVAM, ICCVAM, ZEBET), national organisations, other independent organisations, or commercial sponsors. The sponsor of a study will normally appoint a validation manager with well-defined roles and responsibilities, to co-ordinate the study (Figure 3). In the case of validation studies performed under the auspices of the OECD the sponsor may be an OECD Expert Group, Task Force, Working Group, or Working Party whose members are nominated by the governments of the respective Member countries. Member countries or stakeholders, such as national and international organisations, industry associations, or non-governmental organisations may also represent the sponsor.

65. The validation manager or management team should ensure that the purpose and objectives of the study are clearly defined, a project plan is developed, and a process for careful oversight of the validation study is in place. The validation management should have collective expertise with the test, in the underlying science and the scientific design, management and evaluation of a validation study.

66. When conducting a validation study of a test method, ordinarily the sponsor would be expected to have safeguards against conflict of interest in place to which the validation study should adhere. If it does not, or they do not cover all aspects of conflict, individuals overseeing the study should not have personal or financial conflicts of interest with the test or the outcome of the validation study. Individuals responsible for validation studies should have expertise with the test, in the underlying science and the scientific design, management and evaluation of a validation study.

67. The individual or team responsible for management of the validation should oversee the validation and serve as the central contact point for co-ordination of the testing by participating laboratories and for receipt of data from those laboratories.

68. The validation manager or management team may delegate essential tasks, such as the statistical analysis, the selection of test chemicals and selection of participating laboratories, including decisions regarding the extent of GLP compliance to be used, to task groups or individuals.

69. The validation manager should review and ensure the integrity of all information concerning the test. Figure 3 demonstrates the role of the validation manager/management team in the validation process and it summarizes essential tasks and interactions for managing validation studies. Schemes describing various approaches for the organisation and management structure of validation studies have been described elsewhere (5)(35)(37)(38)(39)(40)(41). Sponsors should ensure that all information and data generated during validation of the test method are archived and are available for independent review.

Statistical Expertise

70. Sufficient statistical expertise should be available to ensure appropriate design of the validation studies and evaluation of resulting data. A statistical advisor (biostatistician) should be a member of, or a consultant to, the validation manager/management team and be involved in all phases of the validation of new and revised tests. The statistical advisor should be familiar with the biological basis and the practical limitations of the proposed test and of the reference test or endpoint in the ultimate species of interest. This knowledge will aid in the selection of appropriate statistical procedures, and the development of appropriate decision criteria, and can help with communicating the results of the study.

71. The biostatistician should be involved in the design of all phases of validation studies and in the establishment of appropriate procedures for record keeping, data collection, and data submission. The

statistical methods to be used should be defined before the start of the studies and be used to support the proposed study design. However, because the specifics of the data can not always be anticipated, it may be necessary to revise the statistical procedures or identify or develop new procedures or models after receipt of the test data. The statistical evaluation should convey the reproducibility of the study by assessing variability within and among laboratories. Acceptance criteria should be given at the beginning of the study as to the degree of variability and agreement judged acceptable for a valid test.

72. At various stages in the validation process, statistical assessments should be made as to how well results from the new or revised test agree with the reference data using the decision criteria. In addition, the test conditions should be evaluated statistically to provide recommendations for improving the efficiency of the protocol. Such an analysis can also serve to optimise and reduce animal use, for example, by showing the power of the test as a function of the number of animals per group in relation to the reproducibility/variability and response of the endpoint that is being measured.

Conduct of the Prevalidation

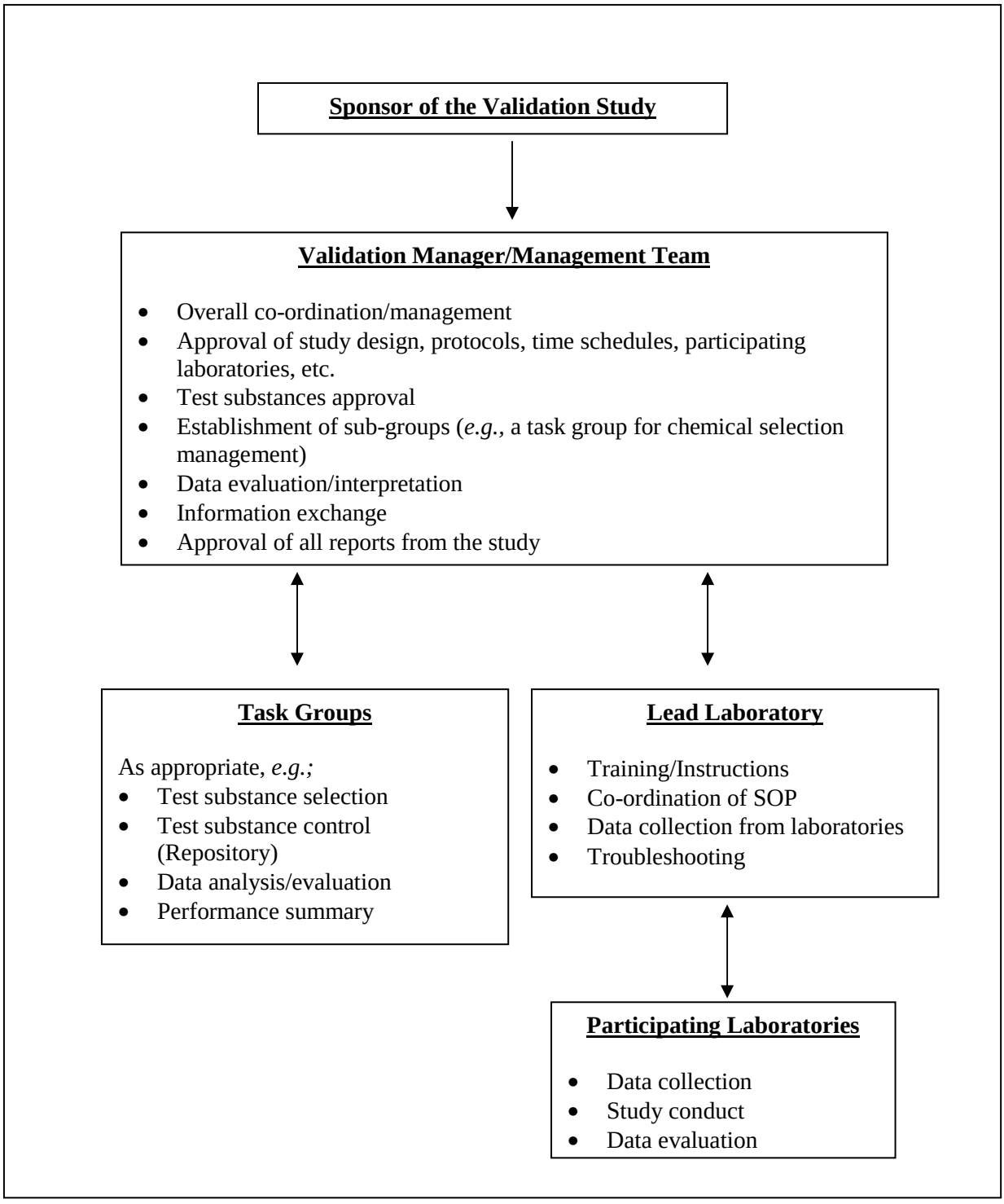
73. The prevalidation could be performed by one or more laboratories experienced in performing tests similar to the new or revised test. The exact design of the study, the number of participating laboratories, the laboratory qualifications required, the number of substances tested, and the number of animals to be used, where applicable, may depend on the specific test, its proposed use and/or biometrical considerations.

74. The prevalidation usually includes participation by one or more laboratories and/or experts in addition to those who were responsible for developing the test. The extent of activity at this phase will vary greatly depending on the specific test and the amount of information already available on its performance, and it may involve a single stage or multiple stages of experimentation, each building on the previous phase of work.

75. Additional activities at the prevalidation stage may include an assessment of the transferability of the test protocol to laboratories inexperienced in the test or the necessary techniques and to resolve questions or inconsistencies arising in this phase. This process may lead to the development of an optimised test protocol and accompanying SOPs, as appropriate, and may provide for development of additional data to be used for refining the evaluation and decision criteria for the test.

76. For certain tests, training of the personnel in the participating laboratories and co-ordination of scoring criteria and nomenclature would be an essential part of the study.

FIGURE 3. ROLE AND INTERACTIONS OF THE VALIDATION MANAGER/MANAGEMENT TEAM



Project Plan

77. The objectives of the validation study should be specifically defined and presented such that they represent a clear, unambiguous and achievable goal. The validation management established for the project is responsible for the definition of the purpose and objectives of the validation. A project plan should be produced so that all involved parties will have a clear understanding of the validation work to be performed. The project plan will form the basis of an agreement among the sponsors, the validation management group, and the lead and participating laboratories. The objectives and specific details of the project and the protocol(s) to be used should be available to all study participants, who should agree, in advance, to carefully follow the project plan and test protocol.

78. In addition to the principles and criteria for a valid test method presented earlier in the text, the following specific indications of readiness should be available before a test is accepted for inclusion in a formal inter-laboratory validation study:

- Definition of the test including its position in a testing program, its advantages and limitations, and the classes and types of test substances that can and cannot be tested.
- Existence of data evaluation or decision criteria based on the results of the prevalidation, which is appropriate for the data generated by the test and the effect to be predicted, modelled or measured (*e.g.*, PM or DIP, where appropriate).
- The specific, detailed protocol or protocols proposed for the inter-laboratory phase of the validation study, including SOPs (as appropriate), the measurements to be performed and the specific types of data to be provided by the test.
- Evidence of intra-laboratory reproducibility of the test, and an initial assessment (prevalidation) of the inter-laboratory transferability.

79. The project plan should contain essential information for the participating laboratories and describe their duties and responsibilities; it should be updated, as needed, as the planning progresses. Depending on the anticipated size and complexity of the study, essential tasks such as the selection of test chemicals, development of a reference data base, selection of a lead laboratory, development of minimum criteria for, and selection of, the other participating laboratories, and selection of data analysis procedures and the performances of statistical analyses, can be delegated to sub-groups or individual members reporting to the validation manager (figure 3).

Participating Laboratories

80. Laboratories participating in the validation study should meet minimum standards in terms of the availability of competent staff, facilities, safety, animal welfare (where applicable) and quality assurance procedures. Ideally, the participating laboratories should have demonstrated competence with the general test method or similar methods and should provide, as a minimum, information on the adequacy of their data documentation procedures, such as their compliance with Good Laboratory Practices (GLP). It is also important to consider that the choice of an inexperienced laboratory may have an impact upon the results of the validation. Ideally, validation studies should be performed, and data recorded and maintained, in accordance with GLP (42)(43)(44)(45)(46)(47). As a minimum, they should always be conducted in accordance with GLP principles, such as, but not necessarily limited to, use of SOPs, adequate data recording and record keeping. The use of standardised software programs for data collection may be useful in this regard.

81. It is important that all laboratories participating in a validation study have a clear understanding of their specific obligations and responsibilities. These include the requirement for strict adherence to test method protocols, the prescribed timetable for submission of data, clarification of ownership of results, and publication procedures (*e.g.*, clearance, publish jointly or independently). It is a distinct advantage to have such issues clearly addressed in formal agreements or contracts, prior to the start of the study. In addition, an international study may involve many languages and legal systems, as well as different financial practices and contractual systems, all of which can complicate validation management if they are not formally agreed to in advance.

82. The minimum and maximum number of laboratories needed for a comprehensive assessment of the validity of the test method will depend on the type of test, the questions being addressed, the overall amount of testing required of each laboratory, the cost of testing, and animal consumption. In many cases, three or four laboratories per test method may be an adequate number for an assessment of the inter-laboratory reproducibility. However, the assessment of the predictive capacity (performance of the test method) may be conducted in fewer laboratories if the data are adequate. Having too many participating laboratories may increase costs, logistical complexity, and animal numbers without necessarily improving the scientific quality of the outcome. Alternatively, having too few participating laboratories may result in insufficient data for an accurate assessment of the inter-laboratory variability. The decision as to the preferred number of laboratories for each validation study should be based on bio-statistical considerations and the issues enumerated in this paragraph adequate in a statistical sense.

83. A lead laboratory should be appointed, and this laboratory should have demonstrated expertise and experience in the discipline to which the test method belongs. It should also have proven experience in conducting the type of test method under evaluation. Validation managers should oversee the development or refinement of the initial protocol and ensure that appropriate training is provided, as necessary, to personnel from the participating laboratories, so that all laboratories perform the test and evaluate the data according to the same procedures and criteria (figure 3). The study director in each laboratory should ensure clear communication with validation management.

Reference Data

84. Relevance needs to be determined for all types of assays. For replacement tests, there should be an adequate reference data base generated in the original test method against which the new (or revised) test method is compared. This reference data base should cover the expected range of responses (strong, moderate, weak and negative) and the range of chemicals to which the assay will be applied (the domain of applicability). A reference data base of this type should provide a measure of the performance of the test method against a wide range of biological and chemical activities. For higher level *in vivo* assays, the nature of the endpoint, a direct measure of toxicity (*e.g.*, histopathological change, organ weight change, etc.) would permit the option of a less extensive data set to assess relevance. Such data sets usually consist of the minimum number of chemicals necessary to cover the range of responses. Inter-laboratory validation studies, when necessary, of such assays should be designed to generate the requisite data to establish the relevance and reliability of the test method while taking into account practical limitations, such as animal use.

85. The ultimate goal of toxicity testing is to protect human and ecological health and safety. For a test that is being designed to replace an existing test, all efforts should be undertaken to obtain reference data from the existing test and, where feasible, the ultimate species of interest. While it is theoretically possible to directly test the species of interest in ecotoxicological tests, standard surrogate species are generally used because they have been adapted to culturing in the laboratory and it is impossible to test every fish or bird of interest, even if suitable number could be obtained from wild populations.

86. For assessing human health effects, data from humans are most relevant;. Human data may be available from several sources (50) including:

- Epidemiology
- Occupational exposure
- Accidents and cases of poisoning
- Clinical studies
- Ethically approved studies in human volunteers

The use of human data raises a number of issues that needs to be addressed, however, in practice it is normally difficult to obtain high quality human data:

- The special challenges of retrospective data comparison;
- The lack of available human data on chemicals that do not show the effect of interest, so that a determination of that test's false positive rate cannot be made;
- The ethical issue of human data development, including national ethical differences in how to approach the use and acquisition of human data,
- The potential lack of accurate information on chemical exposures.

Consequently, the absence of sufficient human data may mean that animal data are the default reference. For example, if the goal of a study is to replace an animal test that is being used as a surrogate for human effects, and human data are lacking, then data derived using that animal species may be the most relevant available and should be used.

87. Classifications of chemical activities in the reference species should be supported by detailed information. The number of reference chemicals meeting these criteria may be insufficient. In such cases, careful judgements including considering animal welfare should be made regarding the need of developing additional reference data.

88. In ecotoxicology, the replacement test issue and process is currently handled somewhat differently compared with other areas. Since there are so many different species of plants and animals occurring in the environment, and only a few from each group can ever be tested, much work has been expended in trying to develop a minimum of tests for each group of organisms. Once these initial tests have been developed, depending on the "importance" (in the ecological and economic sense) of a group, additional species in the same group can be investigated for development as additional test species. Thus, for an important group of organisms (*e.g.*, fish) there is often a list of acceptable species that can be tested. It is up to the regulatory authority to provide guidance on which species should be tested, and when in the hazard identification/assessment process this testing should be conducted. Replacement tests should be based on the same type of organism, *e.g.*, a crustacean replacing an existing crustacean test, however, these direct replacements seldom occur in ecotoxicology.

Selection of Test Chemicals and Chemical Management

89. The validation management has the overall responsibility for organising the selection of chemicals to be used. In general, a chemical selection sub-group may be established for that purpose (figure 3). Because regulatory authorities are the future users of a new test method undergoing validation, it is advisable to involve experts from these authorities in the selection of the test chemicals. Bio-statistical advice should be sought to determine the minimum number of test chemicals necessary for achieving the objectives of the study. This number may need to be corrected if it does not allow a sufficient

representation of relevant chemical groups. This number might vary depending on whether the predictive capacity (performance of the test method) and intra- and inter-laboratory reproducibility have to be assessed concomitantly or independently. Consideration could also be given to the establishment of a chemical repository to purchase, code, and supply the test chemicals and relevant safety information to the participating laboratories.

90. The chemicals should be selected based on the objectives of the study, and the type of test method undergoing validation (49). Therefore, it is imperative that selection of the chemicals to be tested is consistent with the defined objectives of the study. This means the chemicals need to be relevant to the adverse effect of concern, or mechanism of concern, and also be relevant for the species of concern. Chemicals for which toxicological hazard assessment can be made based on measures other than biological tests (*e.g.*, accepted physico-chemical properties or chemical structure characteristics) should not be selected (or kept at minimum), for example, chemicals corrosive to skin because of their extreme pH values. Chemicals should be chosen according to the availability of relevant and reliable reference data, which ideally provide an unequivocal assessment of their activity in the reference test or in the organism(s) of concern. The performance of the test method being validated will be judged on the basis of its responses to these chemicals, so the spectrum of chemicals and chemical classes that are ultimately included in the study should be relevant to the chemicals for which the validated test will be used. In making these selections, it is important to include a sufficient number of positive and negative chemicals from the reference data. In all cases the criteria and reasoning for selection of test chemicals should be transparent and properly communicated, and included in published reports, if appropriate.

91. The chemicals tested should, where possible, be of the highest available purity, or of known composition. All laboratories in the validation study should use, to the extent possible, the same batch of a test chemical. Ideally, each test chemical should be as close as possible, in terms of chemical composition and physico-chemical properties, to the material employed to generate the reference data. However, it often will not be possible to use the same batch or exact formulation as was used to generate the reference test method data. The chemicals should be readily available for testing (preferably from commercial sources) and in sufficient quantity for use by all participating laboratories, including a back-up sample to cover unexpected losses and, to a limited extent, follow-up investigations that may become necessary. They should be stable under the defined conditions of storage for at least the duration of the validation study. Physico-chemical information regarding stability, solubility, volatility, etc., should be provided, where known, as should storage, use, and disposal recommendations.

92. The number of chemicals that should be used to characterise accuracy and reliability of a test method will vary with the purpose and nature of the test method. For the assessment of inter-laboratory reproducibility a subset of test substances used to assess accuracy might be appropriate, provided that the subset adequately represents an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. In the validation of certain *in vivo* tests, such as carcinogenicity tests and neurotoxicity tests, where the endpoints of interest are directly observed, chemicals used in validation studies should cover the expected range of responses (such as strong, moderate, weak and negative) and a rationale should be provided for the number and properties of the substances used. In addition, the experience gained during test development and prevalidation should be taken into account when selecting test chemicals, so as not to waste time and resources. The number, and types, of test chemicals will depend upon many factors, including;

- The type of test being validated;
- the number of participating laboratories;
- the number of toxicity hazard classes to be predicted;
- the range of toxic effects and potencies being investigated;

- the diversity of chemical classes of interest and structural relationships among the chemicals;
- the range of physical states to be considered;
- practical limitations pertaining to the complexity and duration of *in vivo* tests,
- the availability of relevant and reliable reference data for the chemicals; and
- statistical considerations.

Coding and Distribution of Test Samples

93. Chemicals to be tested should normally be coded in validation studies, unless precluded by scientific, technical or safety considerations. Testing laboratories should ensure that there is compliance with the regulations for the transport or storage of dangerous chemicals in the relevant country. The substances to be tested should be independently coded with a unique code, and packaged in a manner that will not reveal their identities (51). For safety considerations, and to aid in the selection of test article diluents/vehicles, handling, and disposal procedures, sufficient information about the physico-chemical properties of each test chemical should be made available to the participating laboratory and to the laboratory's safety officer. It is preferable to have the test substances coded prior to being sent to the laboratories. A less desirable option is to have the designated safety officer of the testing laboratory or other individual not involved with the validation study apply the codes to the chemicals. In all cases, it is recommended that Safety Data Sheets (SDS) for all chemicals to be studied be kept by the designated safety officer in each participating laboratory. Member countries may have specific regulations regarding provision of SDS.

94. The coded test substances should be transported in suitable containers, preferably from a central repository, and should be labelled in accordance with relevant transport regulations (52). Laboratories should be provided, where appropriate, with the information about each of the test chemicals, including:

a) Visual appearance;

b) physical state;

c) weight or volume of the sample of test chemical dispatched;

d) physico-chemical data, such as pH, volatility, and stability, solubility, chemical reactivity; and

e) storage instructions.

95. If specific chemicals require additional safeguards or control, the relevant information should be included with the physico-chemical information sheet unless this would reveal the chemical's identity, in which case an alternative means should be established to ensure safe use, handling and disposal. Participating laboratories should be provided with a sealed package containing necessary information about the health hazards of the test chemicals. This should include clear instructions for action in the case of accidents, and information needed for decontamination or disposal of contaminated animals and laboratory supplies, and of excess chemicals and working solutions. Advance notification of shipping, information regarding the storage requirements, and safety-related information about the specific chemicals, should be provided.

96. For validation studies that use coded hazardous chemicals, it may be advisable to treat all coded chemicals as if they have the same toxicity as the most hazardous chemical in the group. In such a case, health and safety information provided to the laboratories on this chemical should be considered to apply to all the chemicals in the group. In certain situations, such an approach will not be possible (*e.g.*, the emergency response procedures for one or more of the chemicals differ from those for the most hazardous chemical on the basis of significant differences in water solubility). For those situations, one approach is

to provide a sufficient number of sealed envelopes for each coded chemical separately, containing health and safety data and marked with the code number for that chemical. The sealed envelopes should be available in all working areas where the chemicals are handled, to be available in case of an emergency. However, laboratory staff should be instructed to open and read a given envelope only under such an extreme situation. At the conclusion of the study, the participating laboratory should return all envelopes (sealed and open) and provide written explanations for every opened envelope. In some cases, a specific health and safety plan may need to be developed by the testing laboratory, in which case the health and safety staff, but not the testing staff, would need to know the identity of the test articles. In all cases, information provided to the participating laboratories should be in accordance with national and local legal requirements.

Test Method Decision Criteria and Data Interpretation

97. The outcome of prevalidation studies will determine if a test method is sufficiently developed, to proceed to an inter-laboratory validation study. A rationale and decision criteria for interpretation of test results, classification of a chemical (*e.g.*, positive, negative, equivocal) and a determination of toxicity values and categories should be clearly defined in the standardized optimized protocol used for the validation study. It is noted that the decision criteria for the interpretation of test results, generally an integral part of validation, should be transparent and unambiguous. Depending on the nature of data generated by the test and also by the sort of test used, different approaches may be used to define the rationale and decision criteria. In those cases where a toxic end point of interest is monitored or measured, often in *in vivo* or *ex vivo* systems, clear decision criteria *e.g.*, for positive, negative or equivocal outcomes, should be defined (*e.g.*, in a DIP).

98. When the toxic end point of interest is not measured or monitored directly, an algorithm may be used to convert the result obtained from using the test method into a measurement or prediction of the toxic effect of interest. An algorithm is often available for tests, such as *in vitro* tests, where the results expressed as a value, or a range of values, are related to a toxic effect of interest. However, they are rarely used for ecotoxicity *in vivo* studies. During validation of the 3T3 NRU Phototoxicity test (OECD Test Guideline 432) it was shown that properly developed probability-based PM's can discriminate between treated and untreated controls (53)(54)(55)(56).

Monitoring of Participating Laboratory Performance

99. In a study that uses a large number of test substances, it may be most effective to proceed in a step-wise fashion. Using this approach, only a subset of the substances is tested at first, and a compliance check is performed on each laboratory, which includes a review of the complete data sets. After this review is completed, a decision is made as to whether each laboratory is performing satisfactorily. Laboratories that are not able in this first step to perform the procedure according to the required protocol and SOPs (as appropriate), or maintain the appropriate record keeping procedures, should not be considered for further participation. In this case, decisions must be made about whether the data from such laboratories are sufficiently robust to use further in the validation process.

100. At interim times during the course of the testing, validation management should determine whether the laboratories are following the agreed test protocol and record keeping procedures. This can usually be accomplished through compliance statements from the laboratories. If considered necessary, a compliance check may be performed, which could include a review of the laboratory's records for a limited number of test samples. Additionally, the quality of the laboratory's work (*e.g.*, levels of contamination of samples, performance errors, high reproducibility of data points) may be assessed. Deviations from the test protocol or record keeping procedures, or evidence of poor laboratory practices, should be identified and documented. A decision should be made whether to correct the problems and continue the testing (by

retesting the inadequately-tested samples) or to eliminate the inadequate data from the validation study. Alternatively, data that were not obtained using the required protocols could be identified but not used in subsequent analyses, and the reasons for their rejection documented. As a general principle, all data from all participating laboratories will be used in the analyses and reports prepared by the lead laboratory or the validation management group.

Inter-laboratory Testing

101. Testing in this phase would not normally proceed until the validation manager is satisfied that the test method protocols are clear and unequivocal, and that the tests can be adequately performed by all participating laboratories (for ring-tests, see paragraph 64). Moreover, validation management should confirm that the procedures to be used for collecting, analysing, and reporting the data are adequate. A definitive schedule should be established so that the test articles are supplied to, and received by, all laboratories within a specific time, and that all laboratories complete their testing assignments within a specified time. At the conclusion of the testing, or at agreed-upon intervals, the full data records on completed samples should be submitted to validation management or the data analyst using the appropriate format, as described in the project plan.

102. Laboratories that develop or validate test methods may not all be familiar with GLP or not be organised to perform studies under strict GLP compliance. Components of GLP that should be followed in all cases are that all protocols, experiment-related notes, and data entries should be complete, detailed, accurate, and annotated with the names of the individuals conducting the work, entering the data and the dates of the work and that all staff involved are adequately trained.

Data Collection

103. If the validation study has a large number of participating laboratories, or if there is an anticipated high volume of test data, separate task subgroups may be responsible for aspects of the study relating to the inter-laboratory co-ordination, training, and/or data collection, management and analysis. In certain cases an independent contractor may be designated to receive, process, and analyse the data and may also serve as the lead laboratory, in accordance with the requirements on transparency(figure 3).

104. The following factors should be addressed when planning a validation study and designing the data collection and evaluation procedures.

- Data to be collected, the data recording procedures, and the procedures for performing data quality checks and analysing the data should be explicitly identified.
- For each test method procedure the number of repeat experiments each laboratory will perform, if any, and the number of independent replicate measures of the endpoint within an experiment, if any, should be specified.
- A standard format for data entry to be used by all participating laboratories should be specified. Care is needed to ensure that such forms and formats are not ambiguous; clear directions should be given on how to report the data and the other test-related information.
- The quality assurance and quality control procedures to which the original data obtained and any values derived there from should be identified.

105. Specific computer-based systems may be used to collect, analyse, and report the data, but the use of such a system should be agreed upon as part of the project at the study design stage. The study plan approved by validation management should specify the data that is to be submitted and, on request, all

individual data should be available. Unless otherwise specified in the study plan, all data should be submitted. Where summary or transformed data are submitted, the original data should be available for analysis.

Data Analysis

106. The statistical analysis procedures used should be relevant to the types of data being analysed and the questions being asked of the data. The study plan approved by validation management should specify whether the participating laboratories will conduct analyses of their test data (according to the agreed-upon procedures) or whether independent statistical analyses will be conducted. The latter approach is favoured. The biostatistician(s) should be sufficiently knowledgeable about the test procedures and the design of the validation study to perform the appropriate analyses and to recognise anomalies in the data. If necessary, provision should be made for validation management's biostatistician(s) to directly contact individual laboratories to clarify aspects of their data reports.

107. In general, two types of information are required from a validation study, namely measures of test reliability or reproducibility, and measures of the relevance of the test method. An assessment is made of the qualitative and quantitative reproducibility of the test between the participating laboratories, including measurements of the variability, within each, and between the different laboratories. Where appropriate, an assessment may be performed of the various parts of the protocol that may be most responsible for affecting reproducibility and that contribute to the variability.

108. The individual laboratory data are analyzed to determine the relevance of the test results. Where applicable, this will include assessing the accuracy of the proposed test method for predicting the outcome of the reference test method. Any adjustment of the decision criteria (*e.g.*, PM) during the study should be fully documented and scientifically justified. At the completion of a validation study, there may be situations where the data clearly indicate that the decision criteria need to be refined in order to increase the predictive capacity for the intended use of the test method (57). The important consideration is that any changes in the decision criteria need to be properly documented, and the effects and consequences of the changes need to be clearly described. In terms of ecotoxicological test methods, decision criteria are often used in a different manner from that used in human health testing. In ecotoxicology, decision criteria may not be an integral part of the validation where the test results do not require interpretation.

109. Several different statistical methods may be used for the analysis of data. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Qualitative or categorical data have often been described in terms of the Cooper statistics (58)(59), such as the sensitivity and specificity of the test results. There are many procedures that can be used to evaluate quantitative data. The specific procedure(s) used should be relevant to the types of data being analysed, the comparisons being made with the reference or other data, and the purposes of the study.

110. Following completion of the inter-laboratory phase of the validation, and compilation and analysis of the data, validation management should make a critical assessment of the outcome of the study. All data from each participating laboratory should be assessed and all factors should be taken into account to determine whether the original goals of the study have been met. This assessment should review every aspect of the validation study in a systematic manner.

Reporting

111. At the completion of the study all data and laboratory records should be archived. The goal should be to have archived data and other laboratory records available, upon reasonable notice, in an easily retrievable format for independent assessment.

112. The results of the validation study and final conclusions should be discussed with all participants of the study. Validation management should attempt to reach consensus on the results and conclusions of the validation study before drafting the study report. The draft study report shall be circulated to the lead laboratory and every participating laboratory for review and comment prior to finalisation. Validation management should review all comments received and make revisions as the group deems appropriate.

113. The final study report describing and discussing the validation study, the results obtained and the recommendations of validation management should be submitted to the sponsor. The final report should be subjected to evaluation by an independent panel of peer reviewers (see chapter V; Independent evaluation of a validation study: peer review) prior to submission to regulatory authorities for regulatory acceptance consideration. The report of the independent peer review and the report of the validation study should be made publicly available and, preferably, an article should be prepared for submission to the scientific literature based on the report.

Record-keeping/Data Dissemination

114. It is preferred, but not always strictly required, that validation studies are performed and reported in accordance with the OECD Principles of Good Laboratory Practices (43). The records should contain GLP compliance statements, and indicate if, and when, coded chemicals were used. Aspects of data collection not performed according to GLP should be fully described, along with the potential impact of not using GLP.

115. The data supporting the test being validated, the detailed protocols under which the data were produced, and all other supporting documentation, should be available for review, in accordance with the requirements described in the study plan approved by validation management. If mathematically transformed data or summary conclusions are provided, they should be linked to the original data.

116. Preferably, there should be dissemination of summary information to interested parties on a continuing basis throughout the validation study, as long as this would not compromise the study. Such information could include:

(i) A public announcement of the undertaking of the study, including its objectives, if considered appropriate;

(ii) reports of the progress and outcome of the study; and,

(iii) publication of a report of the study in the peer-reviewed scientific literature.

Following completion of the validation study and peer review, the publication of a summary report of the study, and reports from the participating laboratories, should be encouraged. In addition, the study sponsor should consider how to contribute to the transparency of the validation study. Some validation processes make data available for review by interested parties by depositing such data in a repository open to the public.

V. INDEPENDENT EVALUATION OF A VALIDATION STUDY (PEER REVIEW)

117. Independent scientific evaluations (peer review) of test methods may be conducted to assess the extent that a proposed test method has addressed established validation and acceptance criteria for a specific proposed use, and to provide expert advice on the proposed test method. This independent assessment is normally conducted by a panel of qualified scientists that have not been involved in the development or validation of the test method, and who do not stand to benefit financially or professionally as a result of the outcome of the review. Independent peer review provides a critical assessment of the usefulness and limitations of a test method for its proposed use, and aids regulatory authorities in determining the acceptability of a test method. This process is an important step following the validation of a new or revised toxicology test method. All aspects of all peer reviews should be fully transparent and allow for consideration of stakeholders comments.

Mechanisms for Peer Review

118. There are several options for independent peer review of test methods in the development of OECD Test Guidelines, including:

(i) Peer review co-ordinated by an organisation established or mandated by law to evaluate test methods;

(ii) peer review co-ordinated by a government agency of a Member country; and,

(iii) OECD co-ordinated peer review.

119. Sponsors or other appropriate parties could contact organisations that specialize in test method evaluations, *e.g.*, ECVAM or ICCVAM, to arrange for an independent peer review. These organisations provide a resource for assessing the validation status of test methods and a source of guidance for other organisations and can conduct and assess validation status. .

120. It is important for the sponsor of the test method to carefully document the data available to support the validation review process. Principles and criteria described in table 1 should be met and this information should be provided to the peer reviewers. Accordingly, a submission data package documenting the validation status should include:

- Scientific and regulatory rationale for the proposed test method
- Rationale for essential test method protocol components
- A complete description of the substances used in the validation of the proposed method, and the rationale for their selection
- *In vivo* or other appropriate reference data for test substances used to assess the test method's accuracy (where appropriate)
- All available data and results from the proposed test and reference methods
- Test method performance (accuracy)
- Test method reliability (repeatability/reproducibility)
- Statement assessing test method data quality
- Other pertinent scientific reports and reviews
- Assessment of test method refinement, reduction and replacement

- Evaluation of usefulness and limitations of the test method
- References
- Supporting materials

121. Practical guidance for organizing test method submissions that has been found to be helpful in assuring that peer review panels have sufficient information to evaluate the validation status of test methods has been developed and used effectively by validation agencies such as ICCVAM (60). While this should not be understood as a binding standard for other parties, it should be considered as one way to provide adequate documentation for peer review (see also paragraph 145-146, Criteria for regulatory acceptance).

122. The party that submits a test method to the OECD Test Guideline Programme as the basis for development of a Test Guideline should indicate if peer review has already been conducted or propose a peer review approach to be used for its evaluation. Where peer review has been completed, a full and complete report of the independent panel should be provided, including a detailed rationale for conclusions and recommendations, along with all comments provided on the test method. When a proposed test guideline is submitted to the OECD based on a test method which has not undergone independent peer review, the OECD should discuss and determine an appropriate responsible party to organize the necessary independent review.

123. Flexibility in the peer review process is considered to be essential. However care should be taken to ensure that the peer review continues to be a balanced, expert, and fully open and transparent process.

Selection of Peer Reviewers

124. The experts who serve as peer reviewers of validation studies should be technically competent, credible and able and willing to perform their assessment and writing tasks. In addition, conflict of interest criteria should be described. For example, some validation organisations require that the peer reviewer or any immediate family member should not have financial or other conflict of interest with the test or the outcome of the peer review. In order to ensure the integrity of the peer review process, the organisation overseeing the peer review process should develop and apply specific criteria for assessing experts who are candidates for the review. The following specific selection criteria could be considered:

- The experts have demonstrated expertise in one or more relevant scientific disciplines. The peer review panel should include at least one expert from each of the identified disciplines. For certain essential disciplines, it may be useful to have more than one expert on the review panel.
- It is also important that some peer reviewers collectively have expertise in the design, conduct and evaluation of validation studies for new and revised toxicology tests and, when appropriate, that one or more reviewers have understanding of animal welfare issues and the principle of the 3Rs of animal welfare.
- It is also critical that peer reviewers are independent and not subject to any conflict of interest, either in terms of previous substantial involvement in the process or of material interest in the outcome of the review, *e.g.*, financial. Alternatively, any particular interest in the outcome should be declared at the start of the peer review process and the panel, as a whole, should be balanced and independent.

125. A panel of peer reviewers may be established for each validation project separately, as is done by ICCVAM (60) and ECVAM, the US National Academy of Sciences and, on a case-by-case basis, in the OECD, or an established independent expert panel could be augmented with appropriate expertise. The

mechanism for nomination/selection of the peer reviewers may vary with the organization responsible for organizing the review. Decisions on the mechanism to be used are made on a case-by-case basis. In all cases, nominations of qualified experts should be broadly solicited from appropriate scientific and stakeholder communities. It is crucial that the panel as a whole is balanced and that the peer review process is open and transparent, especially if stakeholders or experts that have been involved in the validation are members of the peer review panel. To allow for wide global acceptance of the validated test method, organisations or countries that have the task of arranging an independent evaluation of the validation process should consider whether the peer review panel should be international in composition.

Charge to Peer Reviewers

126. To facilitate the peer review process, the peer review panel could be provided a list of scientifically directed questions appropriate for the nature of the test method being evaluated, however, the list should not restrict the scope of the peer review. These questions could serve as a template for the peer review panel's evaluation. The questions are usually of a standard nature; *e.g.*, has the test method data been collected using studies designed and conducted to comply with appropriate international standards for GLP? Other questions may be of a more specific nature to assess the appropriateness of the protocol components with respect to attaining the objective of the proposed test method. The organisation overseeing the peer review process should have the responsibility for developing questions to the reviewers to ensure that all criteria are assessed by the panel

127. The charge to the peer reviewers needs to pose searching and insightful questions that will elicit input on the usefulness and limitations of the critical aspects of the test method for its specific proposed use. The reviewers should provide comprehensive reviews on the extent to which each of the validation criteria have been addressed, and their views on the proposed standardized protocol and decision criteria. The panel should identify the chemical and physical properties of chemicals and products for which the test method is likely to be useful and those for which it has not been adequately evaluated, or not shown to be useful. Finally, the panel should comment on the appropriateness of the test performance (*i.e.*, its accuracy and reliability) for its intended purpose. The review should be complete, objective, and credible.

Peer Review Process

128. As described above, the first step of the peer review is provision of a complete test method submission package to the peer reviewers. This package could include the charge, the validation study reports and other documents that are to be evaluated by the reviewers, as well as other relevant documents that are to be considered by the reviewers, in responding to the charge.

129. Whenever possible, all materials reviewed by the peer review panel should be made available to appropriate stakeholders. Stakeholder comments should be provided to the members of the peer review panel for their consideration. In some Member countries there is a requirement to include the public in this commenting process. Ideally, it is desirable the peer review process be open and transparent. Some validation agencies address this by having the peer review panel convene in public sessions. Practical considerations, particularly for international reviews, should determine the best process for the peer review panel. At the end of the review process, the overall assessment of the usefulness and limitations of the test method should be included in the final peer review panel report. This final assessment will be based on the questions posed to the panel.

130. Peer reviews do not necessarily require full consensus agreement on each issue. However, a description of any dissenting views and their supporting rationales should be provided. The results of the deliberation by the peer review panel should be available in a written report subsequent to the final recommendation of validation status of the test method by the panel. The peer review report should not

only discuss the usefulness and limitations of the proposed test but should provide recommendations for future actions concerning the test method.

131. The test method validation sponsor has the ultimate responsibility to approve the recommendations from the peer review panel and decide whether the test method has been validated for its intended purpose and should be recommended for regulatory acceptance. The sponsor of the test method will also be the responsible party to address critical issues during the peer review *e.g.*, how to address views from individual members of the panel and how to deal with the results from under-performing laboratories.

132. Ideally, the peer review report should be available for widespread public dissemination, and the report, or a synopsis of the report, be published in a scientific peer-reviewed journal. Ultimately, each regulatory authority will have to make its own decision whether or not to accept a validated test method is suitable for its purpose.

133. In certain instances, a previously reviewed test method which has been subjected to further revision or development may be submitted for an additional or subsequent review of its validation status. The level of effort devoted to this subsequent validation review should be commensurate with the degree and importance of changes that have occurred to the protocol components of the test method.

VI. INTERNATIONAL REGULATORY ACCEPTANCE OF VALIDATED TESTS

134. To conserve resources and time a test method can undergo validation studies and peer-review evaluation prior to its consideration for an OECD Test Guideline. However, this is currently not a requirement in the process for developing OECD Test Guidelines.

135. In a research and development setting, toxicology test procedures are conducted during the normal course of the research, and the scientist performing the procedure, and those using the data, have sufficient knowledge of and experience with the procedure to make decisions based upon the data. Validated test methods are reliable indicators of toxicological effects in the regulatory environment where test results are used to support the decision-making process.

136. The regulatory acceptance process has generally been on a case-by-case basis, and regulatory authorities have the option to accept results generated using a test method that has not undergone what today would be considered formal validation (*e.g.*, methods used in mechanistic studies that could help underpin or explain results derived from other tests). However, acceptance of a test method by a specific regulatory authority does not necessarily indicate universal acceptance by other authorities. Acceptance policies differ from country to country and even, at times, among regulatory authorities within a country.

137. Validation as described in this Guidance Document contributes strongly to the international acceptance of any proposed test method and encourages and supports worldwide Mutual Acceptance of Data (MAD) (61). However, regulatory authorities may still have additional questions on the test method beyond its established reliability and relevance, which could affect its regulatory acceptance. Harmonisation of international regulatory acceptance of adequately validated test methods may be achieved by considering the guidance provided in this document. The regulatory acceptance of tests that have not been subjected to prevailing validation processes is discouraged. In cases in which validation is not considered necessary or appropriate, a written justification should be available.

Validation Study Outcomes

138. Following adequate validation studies demonstrating the utility and the applicability of a test method, it may be considered for adoption by regulatory agencies. A validated test method may be submitted to the OECD for formal adoption as an OECD Test Guideline.

Criteria for Regulatory Acceptance

139. After a test method has undergone formal validation and is considered acceptable for specific proposed uses, a recommendation may be made that it be adopted as an OECD Test Guideline. As mentioned earlier in this document, regulatory acceptance would be greatly facilitated by the involvement, as early as possible in the validation process, of the regulatory agencies to which test results derived from the validated method will be submitted. In considering the regulatory acceptance of a new test, the principles and criteria described in Table 2 are important.

TABLE 2: PRINCIPLES AND CRITERIA FOR REGULATORY ACCEPTANCE OF A NEW OR UPDATED TEST METHOD

a) **The submitted test method and supporting validation data should have been subjected to a transparent and independent peer review process.**

The peer review should be made by experts in the field, not substantively involved in the development or validation of the test, knowledgeable of the method and data evaluation, and not materially or financially affected by the outcome of the review. Selection of peer reviewers shall be based upon scientific qualifications, expertise and experience, regardless of employer or affiliation.

b) **Data generated by the test method should adequately measure or predict the endpoint of interest. For replacement test methods, the data should show a linkage between the proposed test method and an existing test method, and/or the proposed test method and effects in the target or model species.**

c) **The test method should generate data useful for hazard/risk assessment purposes.**

New test methods may fill a recognised data gap. The test method may be useful alone or as part of a test battery or tiered testing approach. For substitute test methods the generated data should be at least as useful as, and preferably better than, those obtained using existing test methods.

d) **The submitted test method and supporting validation data should adequately cover a spectrum of chemicals and products representative of those administered by the regulatory programme or agency for which the test method is proposed, and the applicability and limitations of the test method should be clearly described.**

e) **The test method should be sufficiently robust (relatively insensitive to minor changes in the protocol) and transferable among properly equipped laboratories with adequately trained staff.**

The test method preferably allows for standardisation. If highly specialised equipment, material, or expertise is required, means should be provided to facilitate test method transferability.

f) **The test method should be time and cost effective and likely to be used in a regulatory context.**

g) **Justification should be provided (scientific, ethical, economical) for the new or updated test method in light of already existing test methods.**

Where appropriate, animal welfare concerns including the principle of the 3Rs of replacement, reduction and refinement, should be addressed.

140. In a number of OECD member countries, a new or revised test can be adopted for use if it meets three criteria, namely;

- It has been sufficiently validated; and,
- it provides at least as much scientifically credible information as an existing test while using fewer or no animals or while causing less animal suffering (7).
- It improves the safety assessments for man and the environment.

In such cases the principles and criteria listed for both validation and regulatory acceptance should be fulfilled.

From Protocol to Test Guideline

141. Based on past experience, regulatory bodies will routinely seek test procedures that provide for both accuracy and flexibility in fulfilling testing requirements for regulatory purposes. This may result in the use of a general international Test Guideline (TG) rather than a specific test method protocol.

142. It is suggested that an expert group with expertise with the test method draft the Test Guideline. The expert group should be able to distinguish between specific elements that it believes are critical to the generation of valid test results and, therefore, should be required in the Test Guideline, and those items that can be left to professional judgment and are, therefore, discretionary. If members of the expert group cannot agree on whether a parameter should be a required or discretionary element, it would be advisable to conduct a study to determine the impact of variations in that parameter on test results. It is recommended that the expert group develop the Test Guideline after the test protocol is optimized, standardized and validated.

VII. NEW TEST SUBMISSIONS: SUPPORTING DOCUMENTATION

143. When a new test method is considered ready for proposal as an OECD Test Guideline, submission to the OECD Secretariat may be done by any of the following processes:

i. A Member country through its National Co-ordinator;

ii. the EU through the EC;

iii. an industry association through BIAC to the OECD;

iv. invited experts via a National Co-ordinator.

144. Details of the process for adopting a test method as an OECD Test Guideline are provided in a separate Guidance Document (4).

145. As a practical example of how to organize information necessary to substantiate the validity of a test method proposed for regulatory acceptance consideration, material from the guideline for "the Nomination and Submission of New, Revised, and Alternative Test Methods" by ICCVAM has been cited (60). This guideline, which is applicable to ICCVAM (the focus of which is the validation of alternative methods) and its 15 member federal agencies, should not be understood as a binding standard for other parties, but serves as an example found to be an effective and efficient way in one OECD Member country to assemble and organize the necessary information and data to facilitate assessment of the validation status of a test method according to established OECD or other validation criteria. Components of such a submission follow.

Introduction and Rationale for the Proposed Test Method

146. For purposes of a new test submission, a description should be provided of how the proposed test method can be used in the context of current or anticipated regulatory applications (*e.g.*, as a screen in a tiered testing strategy, as an adjunct test to provide mechanistic information, as a substitute or replacement for an existing test method). The mechanistic basis of the proposed test method and the context in which it will be used to measure or predict the toxicological activity of a test material or substance should be discussed, as well as what is known about the similarities and differences of modes and mechanisms of action in the test system compared to the species of interest (*e.g.*, humans for human health-related toxicity testing). If applicable, the extent to which the proposed test method meets the performance standards of a mechanistically and functionally similar validated and accepted test method should be addressed. The sponsor should indicate the relevant classes of chemicals and products that can and cannot be evaluated using the proposed test method and any known limitations of the test method. Finally, the sponsor should indicate where and how the proposed test method might be included in the overall safety or hazard assessment process. In particular, if the proposed test method is part of tiered or battery approaches, the weight given to the new method relative to other tests in the tier or battery should be addressed.

Test Method Protocol Components

147. The sponsor should use this section to explain and describe the basis for decisions on critical functional, structural, and procedural elements of the test method protocol (a complete, detailed protocol for the proposed test method should be provided in an appendix to the submission). This would include a description of the extent to which the proposed test method protocol is similar to the protocol of a validated mechanistically and functionally similar test method for which performance standards exist. The basis and impact of any protocol modifications made during the validation of the proposed test method should be discussed. The technical parameters of the proposed test method (*e.g.*, vehicles, exposure time), the nature

of the response evaluated, and the basis for proposed concurrent controls should be described. Concurrent controls (negative, solvent, and positive), as appropriate, serve as an indication that the test method is operative under the test conditions and provide a basis for experiment-to-experiment comparisons; they are usually part of the acceptance criteria for a given experiment. The acceptable ranges for the control responses and historical data used to establish the acceptable range should be included.

148. The nature of the data to be collected, the methods used for data collection, the type of media in which data are stored, measures of variability, the statistical or non-statistical methods used to analyze and evaluate the data (including methods used to explore a dose-response relationship), and the decision criteria (and their rationale) used to classify the response as positive or negative, if applicable, should be described. The criteria for dose selection and the number of animals required, if any, for dose selection and for the actual test should be stated. Both statistical and non-statistical methods used for data evaluation should be justified. Any confidential information associated with the proposed test method should be indicated clearly; however, the inclusion of confidential information is discouraged.

149. The number of replicate and/or repeat experiments needed to ensure an adequate study should be provided, and the basis for the study design should be described. If replicate or repeat experiments are not part of the proposed test method protocol, a rationale should be provided.

150. The basis for selection of the proposed test method system should be provided. If an animal model is used, the rationale for such use should be provided along with the rationale for selecting the species, strain or stock, sex, acceptable age range, diet, frequency of dosing, the number of doses, and other applicable protocol components should be included.

151. If the test method employs proprietary components, the procedures used to ensure their integrity (in terms of reliability and accuracy) from lot-to-lot and over time should be described. Also, procedures that the user may employ to verify the integrity of the proprietary components should be described.

Characterisation and Selection of Substances Used for Validation of the Proposed Test Method

152. Since a test method may be determined to be effective for the evaluation of only certain classes of chemicals rather than all types of chemicals and products, it is important to provide information regarding those chemicals/classes which were used in the validation study. The rationale for the numbers and types of substances tested during the validation process should be described. The specific chemical or formulation names and relevant chemical and product classes for the substances tested should be specified. Additional consideration should also be given to the fact that not all data sets will be homogeneous for a given chemical characteristic (*e.g.*, water solubility). In such cases, it may be useful to separate the data set into smaller, more uniform subsets for data analysis based on physical-chemical properties or chemical class. To the extent possible, the following information should be provided for each test substance:

- Chemical Abstracts Service Registry Number (CASRN)
- Physical and chemical characteristics
- Solubility (*e.g.*, Kow)
- Concentrations tested
- Purity
- Source
- Stability of the test substance in the test medium

153. Any characteristics of the test chemicals that may have direct impact on test method accuracy and/or reliability should be described. Information regarding the use of coded substances and blind testing during the validation process should be included. In the case of mixtures, the constituents and their relative concentrations should be stated whenever possible. For a proposed test method mechanistically and functionally similar to a validated test method with established performance standards, the extent to which the reference chemicals recommended in the performance standards were tested in the proposed test method should be discussed, and any deviations from this list should be justified. In situations where a listed reference chemical is unavailable, the criteria used to select a replacement chemical should be described. To the extent possible, when compared to the original reference chemical, the replacement chemical should be from the same chemical/product class and produce similar effects in the *in vivo* reference test method. In addition, if applicable, the replacement chemical should have been tested in the comparable validated test method. If fewer than the listed reference chemicals are used, a rationale should be provided. In addition to a written description of the substances tested, it is desirable that the information be presented in tabular format, such as that depicted in Table 3. It is preferable that this information be provided in both printed and electronic formats.

Table 3. Characterisation of Substances Tested.

Chemical or Product Name	**CAS-RN**	**Chemical Class**	**Product Class**	**Concentrations Tested**	**Purity**	**Supplier or Source of Substance**	**Physical and Chemical Characteristics**

In Vivo Reference Data Used to Assess the Accuracy of the Proposed Test Method

154. If the proposed test method is intended to replace or substitute for an existing *in vivo* reference test method, then a comparison of data between the proposed test method and the *in vivo* reference test method is necessary. Data available from the target species for the biological or toxicity endpoint of interest should also be provided. The submission should include:

- Comparative data for the same substances tested using the *in vivo* reference test method and, where available and applicable, from human or other target species studies. If possible, individual animal and human data should be provided.
- The criteria used to select the *in vivo* reference test method (or human) data.
- The source of the *in vivo* reference test method data (*e.g.,* the literature citation for published information, the laboratory study director, the year generated for unpublished data).
- A description of the protocol(s) employed to generate the *in vivo* reference test method or human data. Any modifications to the *in vivo* reference test method protocol(s) should be stated clearly for each data set, along with a discussion of the potential impact of these modifications on the assessment of the accuracy of the proposed test method.
- A description of the quality of the *in vivo* reference test method data, including the extent of GLP compliance (45)(64)(65)(66)(67)(68) and the use of coded test chemicals.
- The original study data for the *in vivo* reference test method studies or references as to their availability.

- A summary of the availability and use of other, relevant toxicity information from the species of interest (*e.g.*, data from human studies, accidental exposures for human health-related toxicity test methods, results of post-marketing surveillance).

Test Method Data and Results

155. The data generated by testing substances using the proposed test method protocol are reported in this section. Any protocol modifications made during the development process and their impact should be stated clearly for each data set. All data, both original and derived, should be submitted, along with each laboratory's summary judgment regarding the outcome of each study. The submission should include data (and explanations) from all studies, whether successful or not. The statistical approach used to evaluate the data should be described and justified.

156. It is also important to describe the lot-to-lot consistency of the test chemicals, the time-frame of the various studies, and the laboratory(ies) in which the studies were conducted. A coded designation for each laboratory involved in an intra-laboratory evaluation of test method reliability and accuracy is acceptable. Any original data not submitted should be available for review, if requested.

157. Results should preferably be presented in tabular form for easy comparison of results from the reference test methods with those from the proposed test method. Table 4 depicts a suggested tabular format for presenting the results, and can be used for presenting the information used in the accuracy assessment. This information should be provided in both printed and electronic formats.

Table 4. Test Method Accuracy Assessment.

Chemical or Product Name	**CAS-RN**	**Chemical Class**	**Product Class**	**Result Using Proposed Test Method (quantitative)**	**Result Using Proposed Test Method (+/-)**	**Result Using Reference Test Method (quantitative)***	**Result Using Reference Test Method (+/-)**	**References or Data Sources**	**Comments**

*Where possible, data from the *in vivo* reference test method should be separated into single columns for each species with available information. Human data should be always presented independently of nonhuman data. If applicable, corresponding data obtained using the mechanistically and functionally similar validated test method with established performance standards should be provided.

Test Method Relevance (Accuracy)

158. This section should describe the accuracy (*e.g.*, sensitivity, specificity, positive and negative predictivity, false positive and false negative rates) of the proposed test method and it should be compared to that obtained for the reference test method currently accepted by regulatory agencies and also compared to data or recognized toxicity information from the species of interest (*e.g.*, humans for human-health-related toxicity testing). In instances where the proposed test method is measuring or predicting an endpoint for which there is no pre-existing test method, the predictions should be compared to relevant information from the species of interest in order to determine the accuracy of the responses. In cases where the proposed test method is mechanistically and functionally similar to a validated test method with established performance standards, the accuracy of both test methods should be compared. When the results obtained using the proposed test method is discordant from that obtained using the comparable validated test method, the frequency of correct predictions of each test method compared to available

toxicity information from the species of interest should be presented. The basis for any discordance in results for the following comparisons should be discussed:

- The proposed test method and currently accepted reference test methods.
- The proposed test method and, if applicable, the comparable validated test method with established performance standards.
- The proposed test method and the accepted reference test method in predicting responses in the species of interest, where data from the species of interest is available.

159. The submission should include a discussion of the strengths and limitations of the proposed test method and should describe salient issues of data interpretation.

Test Method Reliability (Repeatability/Reproducibility)

160. An assessment of test method reliability (repeatability and reproducibility) should be provided. This assessment should include discussion of the rationale for the selection of the substances used to evaluate intra- and inter-laboratory reproducibility, and the extent to which those substances represent the range of possible test outcomes. Outlying values should be identified and discussed. A quantitative statistical analysis of the extent of intra- and inter-laboratory variability, such as that described in ASTM Publication Number E691-92 (62) or coefficient-of-variation analysis should be included. Measures of central tendency and variation should be summarized for historical control data (negative, vehicle, and positive, where applicable). In cases where the proposed test method is mechanistically and functionally similar to a validated test method with established performance standards, the reliability of the two test methods should be compared and the potential impact of any differences discussed.

Test Method Data Quality

161. The extent of adherence to national and international GLP guidelines (46)(63)(64)(65)(66)(67) for the data presented in the submission, as well as the results of any data quality audits, should be included in this section. Deviations from GLP guidelines and the impact of any non-compliance detected in audits should be described. Information on the availability of laboratory notebooks and other data retained by the sponsor(s) for external audits should be stated. All data should be supported by relevant documentation in laboratory notebooks.

Other Scientific Reports and Reviews

162. In this section, the submission should discuss all data from other published or unpublished studies conducted using the proposed test method. Comment should be provided on any conclusions arising from an independent peer-review of the results derived from and performance of the proposed test method and/or from other scientific reviews of the proposed test method. The conclusions of such scientific reports or reviews should be compared to the conclusions reached in the submission. Any other ongoing or planned evaluations of the proposed test method should be described. In cases where the proposed test method is mechanistically and functionally similar to a validated test method with established performance standards, the results of studies conducted subsequent to the evaluation should be included, and any impact on the reliability and accuracy of the proposed test method discussed.

Animal Welfare Considerations (Refinement, Reduction and Replacement)

163. A description should be included of how the proposed test method will refine, reduce, or replace animal use as compared to currently employed methods used for the endpoint of interest. If the proposed

test method requires the use of animals, the rationale for such use should be provided. A description of the sources used to determine the possible availability of alternative test methods that would refine, reduce, or replace animal use for the endpoint of interest should be provided (68)(69). The description should include, at a minimum, the databases searched, the search strategy, the search date(s), the database search results, and the rationale for not utilizing available alternative methods. The basis for determining the appropriate number of animals for the proposed test method should be described. If the test involves potential animal pain and distress, the procedures and approaches that have been incorporated to minimize and, whenever possible, to eliminate the occurrence of such pain and distress should be discussed (10).

Practical Considerations

164. In this section, the cost and time involved in conducting a study using the proposed test method should be specified and compared to the reference test method(s) and, if applicable, to the mechanistically and functionally similar validated test method with established performance standards. Also included in this section should be:

- A discussion of the facilities and major fixed equipment needed to conduct the test method.
- The general availability of other necessary equipment and supplies.
- The required level of training, expertise, and demonstrated proficiency needed by the study personnel.

References

165. A listing of all publications referenced in the submission should be provided.

Supporting Materials

166. Appendices should contain:

- A detailed protocol for the proposed test method
- Copies of all relevant publications, including those containing data from the proposed test method, the *in vivo* reference test method, and if applicable, a comparable validated test method
- All available original data used to evaluate the validity of the proposed test method
- Suggested performance standards to be considered, if performance standards for the proposed test method do not exist.

VIII. REFERENCES

1. OECD (1996). Report of the OECD Workshop on "Harmonisation of Validation and Acceptance Criteria for Alternative Toxicological Test Methods" (Solna Report). OECD, Paris, 1996 (60 pp). [ENV/MC/CHEM(96)9]

2. OECD (2002). Report of the OECD Conference on "Validation and Regulatory Acceptance of New and Updated Methods in Hazard Assessment" that was held in Stockholm, Sweden, 6-8 March 2002. Stockholm Conference. [ENV/JM/TG/M(2002)2/ADD1]

3. OECD (2004). Report of the Workshop on Data Interpretation Procedures (DIPs), held in Berlin, Germany 1-2 July, 2004. [ENV/JM/TG/M(2004)2]

4. OECD (1995). Guidance Document for the Development of OECD Guidelines for Testing of Chemicals. Environment Monographs No. 76, OECD, Paris, 1993. [OCDE/GD(95)71] Available: [http://www.oecd.org/document/30/0,2340,en_2649_34377_1916638_1_1_1_1,00.html]

5. Balls, M., Blaauboer, B., Brusick, D., Frazier, J., Lamb, D., Pemberton, M., Reinhardt, C., Roberfroid, M., Rosenkranz, H., Schmid, B., Spielmann, H., Stammati, A.-L., and Walum, E. (1990a). Report and Recommendations of the CAAT/ERGATT Workshop on the "Validation of Toxicity Test Procedures". ATLA 18, 313-337.

6. Balls, M., Botham, P., Cordier, A., Fumero, S., Kayser, D., Koëter, H.B.W.M., Koundakjian, P., Lindquist, L.G., Meyer, O., Pioda, L., Reinhardt, C., Rozemond, H., Smyrniotis, T., Spielmann, H., Van Looy, H., van der Venne, M.-T., and Walum, E. (1990b). Report and Recommendations of an International Workshop on "Promotion of the Regulatory Acceptance of Non-animal Toxicity Test Procedures. ATLA 18, 339-344.

7. OECD (1990) Environment Monographs No. 36. Scientific criteria for validation of *in vitro* toxicity tests.

8. ICCVAM (1997). Validation and regulatory acceptance of toxicological test methods: a report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods. NIH Publication No: 97-3981, March, 1997, National Institute of Environmental Health Sciences, (NIEHS), Research Triangle Park, North Carolina, USA. (105 pp). Available: [http://iccvam.niehs.nih.gov/docs/guidelines/validate.pdf]

9. Russell, W.M.S. and Burch, R.L. (1959). The principles of humane experimental technique. London, UK, Methuen, 238 pp.

10. OECD (2000). Guidance Document on the Recognition, Assessment, and use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluations. OECD Environment Health and Safety Publications Series on Testing and Assessment No. 19, OECD, Paris, 2000, 38 pp. Available: [http://www.oecd.org/document/30/0,2340,en_2649_34377_1916638_1_1_1_1,00.html]

11. Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Janusch Roi, A., Prieto, P., Sabbioni, E., Scott, L., Worth, A. and Zuang, V. (2004). A Modular Approach to the ECVAM Principles on Test Validity. ATLA 32, 467-472. Available at: [http://ecvam.jrc.it:8080/publication/Hartung-1.pdf]

12. Bremer, S., Worth, A.P., Paparella, M., Bigot, K., Kolossov, E., Fleischmann, B., Hescheler, J. and Balls, M. (2001). Establishment of an *in vitro* reporter gene assay for developmental cardiac toxicity. Toxicol. *In Vitro* 15, 215-223.

13. OECD (2002). Test Guideline 429. OECD Guideline for Testing of Chemicals. Skin Sensitisation: Local Lymph Node Assay. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html]

14. OECD (2001). Test Guideline 414. OECD Guideline for Testing of Chemicals. Prenatal developmental Toxicity Study Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html]

15. OECD (1997). Test Guideline 424. OECD Guideline for Testing of Chemicals. Neurotoxicity Study in Rodents. Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html]

16. Balls, M. and Fentem, J. H. (1997). Progress toward the validation of alternative tests. ATLA 25, 33-43.

17. ICCVAM (2004). Recommended Performance Standards for *In Vitro* Test Methods for Skin Corrosion. NIH Publication No. 04-4510. Available at: [http://iccvam.niehs.nih.gov/methods/ps/ps044510.htm].

18. Worth, A.P. and Balls, M.(2001). The importance of the prediction model in the validation of alternative tests. ATLA 29, 135-143.

19. Balls, M. (1997). Defined structural and performance criteria would facilitate the validation and acceptance of alternative test procedures. ATLA 25, 483-484.

20. Liebsch, M., Traue, D., Barrabas, C., Spielmann, H., Uphill, P., Wilkins, S., McPherson, J.P., Wiemann, C., Kaufmann, T., Remmele, M. and Holzhütter, H-G. (2000). The ECVAM prevalidation study on the use of EpiDerm for skin corrosivity testing. ATLA 28, 371-401.

21. OECD (2001). Supplement to Test Guideline 404: Acute Dermal Irritation/Corrosion". Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html]

22. Worth, A.P., Fentem, J.H., Balls, M., Botham, P.A., Curren, R.D., Earl, L.K., Esdaile, D.J. and Liebsch, M. (1998). An evaluation of the proposed OECD testing strategy for skin corrosion. ATLA 26, 709-720.

23. Worth, A.P. (2004). The Tiered Approach to Toxicity Assessment Based on the Integrated Use of Alternative (Non-animal) Tests. In Predicting Chemical Toxicity and Fate (M.T.D Cronin & D.J. Livingstone, eds). pp. 389- 410. CRC Press; New York, London, Boca Raton.

24. OECD (2001): Supplement to Test Guidelines 405: "Acute Eye Irritation/Corrosion". Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html]

25. Worth, A.P. and Fentem, J.H. (1999). A General Approach for Evaluating Stepwise Testing Strategies. ATLA 27, 161-177.

26. Dearden, J.C., Barratt, M.D., Benigni, R., Bristol, D.W., Combes, R.D., Cronin, M.T.D., Judson, P.N., Payne, M.P., Richard, A.M., Tichy, M., Worth, A.P. and Yourick, J.J. (1997). The development and validation of expert systems for predicting toxicity. The report and recommendations of an ECVAM/ECB workshop (ECVAM workshop 24). ATLA 25, 223-252. Available at: [http://ecvam.jrc.it:8080/publication/index24.html]

27. Worth, A.P., Barratt, M.D. and Houston, J.B. (1998). The validation of computational prediction techniques. ATLA 26, 241-247.

28. Jaworska JS., Comber M, Auer C and van Leeuwen CJ (2003). Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. Environmental Health Perspectives 111, 1358-1360.

29. Eriksson, L., Jaworska, J.S., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environmental Health Perspectives 111, 1361-1375.

30. Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. and Worth, A.P .(2003). Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environmental Health Perspectives 111, 1376-1390.

31. Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003). Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environmental Health Perspectives 111, 1391-1401.

32. CEFIC (2002) ICCA Workshop on "(Q)SARS for human health and the environment: workshop on regulatory acceptance". Held in Setubal, Portugal, March 4-6, 2002. CEFIC, ACC, ECETOC 2002.

33. Curren, R.D., Southee, J.A., Spielmann, H., Liebsch, M., Fentem, J.H., and Balls, M. (1995). The role of prevalidation in the development, validation and acceptance of alternative methods ATLA 23, 211-217.

34. Zuang, V., Balls, M., Botham, P.A., Coquette, A., Corsini, E., Curren, R.D., Elliott, G.R., Fentem, J.H., Heylings, J.R., Liebsch, M., Medina, J., Roguet, R., Van De Sandt, J.J.M., Wiemann, C. and Worth, A.P., 2002 Follow up to the ECVAM prevalidation study on *in vitro* tests for acute skin irritation, ECVAM Skin Irritation Task Force Report 2. ATLA 30, 109-129. Available at: [http://ecvam.jrc.it:8080/publication/index5006.html]

35. Balls, M., Blaauboer, B.J., Fentem, J.H., Bruner, L., Combes, R.D., Ekwall, B., Fielder, R.J., Guillouzo, A., Lewis, R.W., Lovell, D.P., Reinhardt, C.A., Repetto, G., Sladowski, D., Spielmann, H., and Zucco, F. (1995). Practical aspects of the validation of toxicity test procedures. The Report and Recommendations of ECVAM Workshop 5. ATLA 23, 129-147. Available at: [http://ecvam.jrc.it:8080/publication/Workshopreport5.pdf]

36. Fentem, J.H., Briggs, D., Chesné, C., Elliott, G.R., Harbell, J.W., Heylings, J.R., Portes, P., Roguet, R., Van De Sandt, J.J.M. and Botham, P.A., 2001 A prevalidation study on *in vitro* tests for acute skin irritation: results and evaluation by the Management Team. Toxicol. *In Vitro* 15, 57-93.

37. Goldberg, A.M., Frazier, J.M., Brusick, D., Dickens, M.S., Flint, O., Gettings, S.D., Hill, R.H., Lipnick, R.L., Renskers, K.J., Bradlaw, J.A., Scala, R.A., Veronesi, B., Green, S., Wilcox, N.L. and Curren, R.D. (1993). Framework for validation and implementation of *in vitro* toxicity Tests: report of the

validation and technology transfer committee of the John Hopkins Center for Alternatives to Animal Testing. J. Am. Coll. Toxicol. 12, 23-30.

38. Ohno, Y., Kaneko, T., Kobyashi, T., Inoue, T., Kuroiwa, Y., Yoshida, T., Momma, J., Hayashi, M., Akiyama, J., Atsumi, T., Chiba, K., Endo, T., Fujii, A., Kakishima, H., Masamoto, K., Masuda, M., Matsukawa, S., Ohkoshi, K., Okada, J., Sakamoto, K., Takano, K., and Takanaka, A. (1994). First-phase validation of the *in vitro* Eye Irritation Tests for Cosmetic Ingredients. *In Vitro* Toxicol. 7, 89-101.

39. Balls, M and Fentem, J.H. (1999). The validation and acceptance of alternatives to animal testing. Toxicol. *In Vitro* 13, 837-846.

40. Smrchek, J.C. and Zeeman, M. 1998. Assessing risks to ecological systems from chemicals. In: Handbook of Environmental Risk Assessment and Management, P. Calow (ed.), Chapter 3, pp. 24-90. Blackwell Science Ltd., Oxford, UK.

41. Smrchek, J.C., Clements, R., Morcock, R. and Rabert, W. 1993. Assessing ecological hazard under TSCA: methods and evaluation of data. In: Environmental Toxicology and Risk Assessment, STP 1179, W.G. Landis, J.S. Hughes, and M.A. Lewis (eds.), pp. 40-55. ASTM International, West Conshocken, PA.

42. OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring:
- Number 1: The OECD Principles of Good Laboratory Practice (1982; revised 1997)
- Number 2: Revised Guides for Compliance Monitoring Procedures for Good Laboratory Practice
- Number 3: Revised Guidance for the Conduct of Laboratory Inspections and Study Audits
- Number 4: Quality Assurance and GLP (revised 1999)
- Number 5: Compliance of Laboratory Suppliers with GLP Principles (revised 1999)
- Number 6: The Application of the GLP Principles to Field Studies (revised 1999)
- Number 7: The Application of the GLP Principles to Short-term Studies (revised 1999)
- Number 8: The Role and Responsibilities of the Study Director in GLP Studies (revised in 1999)
- Number 9: Guidance for the Preparation of the GLP Inspection Reports
- Number 10: The Application of the Principles of GLP to Computerised Systems
- Number 11: Advisory Document of the Panel on Good Laboratory Practice: The Role and Responsibilities of the Sponsor in the Application of Principles of GLP (1988)
- Number 12: Advisory Document of the Panel on Good Laboratory Practice: Requesting and Carrying Out Inspections and Study Audits in Another Country Available: [http://www.oecd.org/document/63/0,2340,en_2649_34381_2346175_1_1_1_1,00.html]

43. FDA Good Laboratory Practice for Nonclinical Laboratory Studies. 21 CFR Part 58. Washington, DC: U.S. Government Printing Office, 1999. Available: [http://www.gpoaccess.gov/nara/index.html]

44. EPA Federal, Insecticide, and Rodenticide Act (FIFRA); Good Laboratory Practice Standards; Final Rule. 40 CFR Part 160. Washington, DC: U.S. Government Printing Office, 1998. Available: [http://www.gpoaccess.gov/nara/index.html]

45. EPA Toxic Substances Control Act (TSCA); Good Laboratory Practice Standards; Final Rule. 40 CFR Part 792. Washington, DC: U.S. Government Printing Office, 1998. Available: [http://www.gpoaccess.gov/nara/index.html]

46. Japanese Good Laboratory Practice Standards, including:
(1) Pharmaceutical GLP Standard. 1997. Pharmaceutical Affairs Bureau, Ministry of Health and Welfare, Available: [http://www.jsqa.com/en/glp/glp21e.html]; and
(2) Good Laboratory Practice Standards. 1984. Japanese Ministry of Agriculture, Forestry, and Fisheries. Available: [http://www.ovpr.uga.edu/qau/maffintr.html]

47. Cooper-Hannan, R., Harbell, J.W., Coecke, S., Balls, M., Bowe, G., Cervinka, M., Clothier, R., Hermann, F., Klahm, L.K., de Lange, J., Liebsch, M., and Vanparys, P. (1999). The principles of good laboratory practice: application to *in vitro* toxicology studies. The Report and Recommendations of ECVAM Workshop 37. ATLA 27, 539-577. Available at: [http://ecvam.jrc.it:8080/publication/index37.html]

48. Fentem, J.H., Archer, G.E.B., Balls, M., Botham, P.A., Curren, R.D., Earl, L.K., Esdaile, D.J., Holzhütter, H.-G. and Liebsch, M. (1998). The ECVAM international validation study on *in vitro* tests for skin corrosivity. 2. Results and evaluation by the Management Team. Toxicology *in vitro* 12, 483-524.

49. Barratt, M.D., Brantom,, P.G., Fentem, J.H., Gerner, I., Walker, A.P. and Worth, A.P. (1998), The ECVAM International Validation Study on *In Vitro* Tests for Skin Corrosivity. 1. Selection and Distribution of the Test Chemicals. Toxicology *in vitro* 12, 471-482.

50. CETOC (2004) Workshop on the use of human data in risk assessment. Held on 23-24 February, 2004 in Cardiff, UK. Workshop Report No. 3. Available: [www.ecetoc.org]

51. Brantom, P.G., Aspin, P., and Thompson, C. (1995). Supply, Coding and Distribution of Samples for Validation Studies. ATLA, 23, 348-351.

52. Recommendations on the Transport of Dangerous Goods - Model Regulations. United Nations. 12th Revised Edition. (2001). ST/SG/AC.10/1/Rev.12.

53. Spielmann, H., Balls, M., Dupuis, J., Pape, W.J.W., de Silva O, Holzhütter, H-G., Gerberick, F., Liebsch, M., Lovell, W.W., and Pfannenbecker, U. (1998). A study on UV filter chemicals from annex VII of the European Union Directive 76/768/EEC in the *In vitro* 3T3 NRU Phototoxicity Test. ATLA 26, 679-708.

54. Spielmann, H., Balls, M., Dupuis, J., Pape, W.J.W., Pechovitch, G., de Silva, O., Holzhütter, H-G., Clothier, R., Desolle, P., Gerberick, F., Liebsch, M., Lovell, W.W., Maurer, T., Pfannenbecker, U., Polthast, J.M., Csato, M., Sladowski, D., Steiling, W., and Brantom, P. (1998). The International EU/COLIPA *In Vitro* Phototoxicity validation study: results of phase II (blind trial). Part 1: The 3T3 NRU Phototoxicity Test. Toxicol.*In vitro* 12, 305-327.

55. Peters, B. and Holzhütter, H.-G. (2002). *In vitro* phototoxicity testing: development and validation of a new concentration response analysis software and biostatistical analyses related to the use of various prediction models. ATLA, 30, 415-432.

56. OECD (2004). Test Guideline 432. OECD Guideline for Testing of Chemicals. *In Vitro* 3T3 NRU phototoxicity test.
Available: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html]

57. Genschow, E., Spielmann, H., Scholz, G., Seiler, A., Brown, N., Piersma, A., Brady, M., Clemann, N., Huuskonen, H., Paillard, F., Bremer, S., and Becker, K. (2002). The ECVAM international

validation study on *in vitro* embryotoxicity tests: results of the definitive phase and evaluation of prediction models. European Centre for the Validation of Alternative Methods. Altern. Lab. Anim. 30, 151-76.

58. Cooper, J.A., Saracci, R. and Cole, P. (1979). Describing the validity of carcinogen screening tests. Br. J. Canc. 39, 87-89.

59. Worth, A.P. and Cronin, M.T.D. (2001). The use of bootstrap resampling to assess the uncertainty of Cooper statistics. ATLA 29, 447-459.

60. ICCVAM (2003). ICCVAM Guidelines for the Nomination and Submission of New, Revised, and Alternative Test Methods. NIH Publication No: 03-4508. Available at [http://iccvam.niehs.nih.gov/docs/guidelines/subguide.htm].

61. OECD (1981). Council Decision concerning the Mutual Acceptance of Data (MAD) in the assessment of chemicals. OECD, Paris, 1981 [C(81)30] Available: [http://www.oecd.org/dataoecd/39/15/2017640.pdf]

62. ASTM (1992). Standard practice for conducting an intra-laboratory study to determine the precision of a test method. ASTM E691-92. In: Annual Book of ASTM Standards Philadelphia, PA: American Society for Testing and Materials.

63. U.S. EPA. (1998a). Federal Insecticide, Fungicide, and Rodenticide Act: Good Laboratory Practice Standards; Final Rule. 40 CFR Part 160. Available: [http://www.gpoaccess.gov/nara/index.html].

64. U.S. EPA. (1998b). Toxic Substances Control Act (TSCA): Good Laboratory Practice Standards; Final Rule. 40 CFR Part 792. Available: [http://www.gpoaccess.gov/nara/index.html]

65. U.S. FDA. (1999). Good Laboratory Practice for Nonclinical Laboratory Studies. 21 CFR Part 58. Available: [http://www.gpoaccess.gov/nara/index.html]

66. OECD (1998). Principles of Good Laboratory Practice and Compliance Monitoring. ENV/MC/CHEM (98)17. Paris, France: Organisation for Economic Co-operation and Development. Available: [http://www.oecd.org/]

67. Cooper-Hannan, R, Harbell, J, and Coecke, S, (1999). The principles of good laboratory practice: application to *in vitro* toxicology studies. ATLA 27:539-577.

68. USDA (1966). Animal Welfare Act and Regulations. 7 USC: 2131-2156. Available: [http://warp.nal.usda.gov:80/awic/legislat/usdaleg1.htm]

69. Public Health Service. (1996). Public Health Service Policy on Humane Care and use of Laboratory Animals. Washington, DC: U.S. Department of Health and Human Services. Available: [http://grants1.nih.gov/grants/olaw/references/phspol.htm]

ANNEX I

DEFINITIONS AND GLOSSARY

The definitions presented here are to be considered working definitions for the purpose of this document also giving a broader view of the variability of terminology used by different validation agencies and organisations. All definitions and terminology will be consistent with the proposed OECD/IPCS Harmonisation of Terminology.

3Rs (principle of): Reduction (of animal use); Refinement (to lessen pain or distress and to enhance animal well-being); and Replacement (of an animal test with one that uses non-animal systems or phylo-genetically lower species).

Accuracy: The closeness of agreement between test method results and accepted reference values. It is a measure of test method performance and one aspect of relevance. The term is often used interchangeably with "concordance" to mean the proportion of correct outcomes of a test method.

Adjunct test: Test that provides data that add to or help interpret the results of other tests and provide information useful for the risk assessment process

Alternative test: A test that: reduces the numbers of animals required; refines procedures to lessen or eliminate pain or distress to animals, or enhance animal well-being; or replaces animals with non-animal systems or with non-sentient species.

Applicability domain: A description of the physicochemical or other properties of the substances for which a test method is applicable for use.

Assay: Uses interchangeably with Test.

"Catch-up" validation study: A validation study for a test method that is structurally and functionally similar to a previously validated and accepted reference test method. The candidate test method should incorporate the essential test method components included in performance standards developed for the reference test method, and should have comparable performance when evaluated using the reference chemicals provided in the performance standards.

Coded chemicals: Chemicals that are labelled by code when delivered to the laboratory for testing so that they can be tested without the laboratory personnel having knowledge of their identity or anticipation of the test results. Coded chemicals are used to avoid intentional or unintentional bias when performing laboratory tests or evaluating test results.

Concordance: This is a measure of test method performance for test methods that give a categorical result, and is one aspect of "relevance". The term is sometimes used interchangeably with "accuracy", and is defined as the proportion of all chemicals tested that are correctly classified as positive or negative. Concordance is highly dependent on the prevalence of positives in the types of substances being examined.

Data interpretation procedure (DIP): An interpretation procedure used to determine how well the results from the test predict or model the biological effect of interest. See Prediction Model.

Decision Criteria: The criteria in a test method protocol that describe how the test method results are used for decisions on classification or other effects measured or predicted by the test method.

Definitive test: A test that is considered to generate sufficient data to determine the specific hazard or lack of hazard of the substance without the need for further testing, and which may therefore be used to make decisions pertaining to hazard or safety of the substance.

Dose response assessment: The second of four steps in risk assessment consisting in the analysis of the relationship between the total amount of an agent administered to, taken up or absorbed by an organism, system or (sub)population and the changes developed in that organism, system or (sub)population in reaction to that agent, and inferences derived from such an analysis with respect to the entire population.

Endpoint: The biological or chemical process, response, or effect, assessed by a test.

Expert system: a computer-based tool that generates predictions of endpoints by applying (Q)SARs and/or rules designed to recreate the reasoning of experts. Expert systems may also contain a database of experimental data which may be consulted directly and which may be used during the application of the rules.

Exposure assessment: Analysis of the exposure to an organism, system or (sub)population of an agent (including its derivatives). Exposure Assessment is the third step in the process of Risk Assessment.

False negative: A substance incorrectly identified as negative or non-active by a test method, when in fact it is negative or non-active.

False negative rate: The proportion of all positive substances falsely identified by a test method as negative. It is one indicator of test method performance.

False positive: A substance incorrectly identified as positive or active by a test, when in fact it is negative or non-active.
False positive rate: The proportion of all negative (non-active) substances that are falsely identified as positive. It is one indicator of test performance.

Good Laboratory Practices (GLP): Regulations promulgated by a number of countries and national regulatory bodies that describe record keeping and quality assurance procedures for laboratory records that will be the basis for data submissions to regulatory authorities. Also the subject of the OECD Series on "Principles of Good Laboratory Practise and Compliance Monitoring" (42).

Hazard: The potential for an adverse health or ecological effect. The adverse effect is manifested only if there is an exposure of sufficient level.

Hazard identification: The identification of the type and nature of adverse effects that an agent has as inherent capacity to cause in an organism, system or (sub)population. Hazard identification is the first step in the process of Risk Assessment

Hierarchical (tiered) testing approach: An approach where a series of tests to measure or elucidate a particular effect are used in an ordered sequence. In a typical hierarchical testing approach, one or a few tests are initially used; the results from these tests determine which (if any) subsequent tests are to be used. For a particular chemical, a weigh-of-evidence decision regarding hazard could be made at any stage (tier) in the testing strategy, in which case there would be no need to proceed to subsequent tiers.

In silico models: Approaches for the assessment of chemicals based on the use computer-based estimations or simulations. Examples include structure-activity relationships (SAR), quantitative structure-activity relationships (QSARs), and expert systems.

Inter-laboratory reproducibility: A measure of the extent to which different qualified laboratories, using the same protocol and testing the same substances, can produce qualitatively and quantitatively similar results. Inter-laboratory reproducibility is determined during the prevalidation and validation processes, and indicates the extent to which a test can be successfully transferred between laboratories, also referred to as between-laboratory reproducibility.

Intra-laboratory repeatability: The closeness of agreement between test results obtained within a single laboratory when the procedure is performed on the same substance under identical conditions within a given time period.

Intra-laboratory reproducibility: A determination of the extent that qualified people within the same laboratory can successfully replicate results using a specific protocol at different times. Also referred to as within-laboratory reproducibility.

Lead laboratory: The laboratory selected to perform the initial development of a standardised and optimised test method protocol and to train the other laboratory personnel in the protocol procedures for the performance of an inter-laboratory validation study. This laboratory may also be used to produce the reference data against which the performance of the other laboratories will be evaluated.

"Me-too test": A colloquial expression for a test method that is structurally and functionally similar to a validated and accepted reference test method. Such a test method would be a candidate for catch-up validation.

Mechanistic (based) test: A test that provides a direct relationship between the biological effects observed and the biological effects of interest.

Modular approach: A general conceptual framework that combines the use of retrospective and prospective approaches to validation. According to the ECVAM proposal, the information needed to support the validity of a test methods should be organised into seven modules.

Optimised test protocol: A test protocol that has been revised and improved based on the results obtained in prevalidation and validation studies.

Partial replacement test: A test method that enables animal reduction by replacing an animal test for one or more of its endpoints (but not all of them), a limited range of chemicals (but not all chemicals), or a limited range of the response values. A partial replacement test can contribute to the complete replacement of an (animal) test when used in a complementary fashion in a (tiered) testing strategy.

Peer involvement: The interaction of outside experts of equivalent expertise and experience with the experts performing the work during the development of the scientific product. Such interaction enables an open exchange of ideas, data, and insights. Peer involvement can occur throughout the development of the scientific product and adds to the scientific credibility of a product.

Peer review: A documented critical review of a specific scientific work product or products, which is conducted by experts who are independent of the experts who performed the original work but who are collectively comparable in technical expertise.

Performance characteristics: The operational characteristics of a test are the measures of its performance under specific conditions, including its reliability and accuracy, and are an indication of the test's usefulness, limitations, and relevance

Performance standards: Standards, based on a validated test method, that provide a basis for evaluating the comparability of a proposed test method that is mechanistically and functionally similar. Included are (1) essential test method components; (2) a minimum list of reference chemicals selected from among the chemicals used to demonstrate the acceptable performance of the validated test method; and (3) the comparable levels of accuracy and reliability, based on what was obtained for the validated test method, that the proposed test method should demonstrate when evaluated using the minimum list of reference chemicals.

Prevalence: The proportion of positive or negative substances in the types of substances of interest. In practice, prevalence cannot be determined precisely, but can be estimated from a sample of substances taken from the types of substances of interest.

Prevalidation: The initial phase(s) of a validation study. A small-scale study intended to obtain preliminary information on the relevance and reliability of a test method. Based on the outcome of those studies, the test method protocol may be modified or optimised to increase intra- and/or inter-laboratory reproducibility and accuracy in subsequent validation studies

Prediction Model (PM): a formula or algorithm (*e.g.*, formula, rule or set of rules) used to convert the results generated by a test method into a prediction of the (toxic) effect of interest. Also referred to as decision criteria. A prediction model contains four elements: (1) a definition of the specific purpose(s) for which the test method is to be used; (2) specifications of all possible results that may be obtained, (3) an algorithm that converts each study result into a prediction of the (toxic) effect of interest, and (4) specifications as to the accuracy of the prediction model (*e.g.*, sensitivity, specificity, and false positive and false negative rates). Prediction models are generally not used in *in vivo* ecotoxicological tests.

Predictivity (negative): The proportion of correct negative responses among substances testing negative by a test method. It is one indicator of test method accuracy. Negative predictivity is a function of the sensitivity of the test method and the prevalence of negatives among the substances tested.

Predictivity (positive): The proportion of correct positive responses among materials testing positive by a test method. It is one indicator of test method accuracy. Positive predictivity is a function of the sensitivity of the test method and the prevalence of positives among the substances tested.

Proprietary test method: A test method for which manufacture and distribution is restricted by patents, copyrights, trademarks, etc.

Prospective validation: An approach to validation when some or all information necessary to assess the validity of a test are not available, and therefore new experimental work is required.

Protocol: The detailed, unambiguous step-by-step description of a test method that directs the laboratory as to how to perform the test method. The test method protocol includes the listing and description of all preparations, reagents, supplies, and equipment needed, and all criteria and procedures for generating and evaluating test data.

(Q)SARs (Quantitative Structure-Activity Relationships): Theoretical models for making predictions of physicochemical properties, environmental fate parameters, or biological effects (including toxic effects in environmental and mammalian species). They can be divided into two major types, QSARs and SARs. QSARs are quantitative models yielding a continuous or categorical result while SARs are qualitative relationships in the form of structural alerts that incorporate molecular substructures or fragments related to the presence or absence of activity.

Quality assurance: A management process by which adherence to laboratory testing standards, requirements, and record keeping procedures, and the accuracy of data transfer, are assessed by individuals who are independent from those performing the testing.

Reference chemicals: Chemicals selected for use in the validation process, for which responses in the *in vitro* or *in vivo* reference test system or the species of interest are already known. These chemicals should be representative of the classes of chemicals for which the test method is expected to be used, and should represent the full range of responses that may be expected from the chemicals for which it may be used, from strong, to weak, to negative. Different sets of reference chemicals may be required for the different stages of the validation process, and for different test methods and test uses.

Reference data: An agreed-upon set of values against which the values obtained using the new test will be compared.

Reference species: The species used in the reference or traditional test to which a new or revised test is being compared. This may be the species of interest, or it may be a surrogate species when it is not possible to perform testing on the species of interest (*e.g.*, humans).

Reference test method: A test method against which the results from the new test method are being compared.

Regulatory acceptance: The formal acceptance of a test method by regulatory authorities indicating that the test method may be used to provide information to meet a specific regulatory requirement.

Relevance: Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test method.

Reliability: Measures of the extent that a test method can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability.

Repeatability: The agreement among test results obtained within a single laboratory when the procedure is performed on the same substance under identical conditions. (see Reliability)

Replacement test: A test which is designed to substitute for a test that is in routine use and accepted for hazard identification and/or risk assessment, and which has been determined to provide equivalent or improved protection of human or animal health or the environment, as applicable, compared to t the accepted test, for all possible testing situations and substances.

Reproducibility: The agreement among results obtained from testing the same substance using the same test protocol. (see reliability)

Retrospective validation: An assessment of the validation status of a test method carried out by considering all available information and data generated from one or more validation studies.

Ring test: A multi-laboratory validation study in which all laboratories test the same substances using identical test protocols, often used for eco-toxicity test method validation. The purpose of the study is to determine inter- and intra-laboratory reproducibility of a test method. Sometimes used interchangeably with “round-robin test”.

Risk: The probability or degree of concern that a defined exposure to a substance will cause an adverse effect in the species of interest or in the environment. Risk is determined by assessing hazard identification, dose-response, and exposure information.

Risk assessment: A process intended to calculate or estimate the risk to a given target organism, system or (sub)population , including the identification of attendant uncertainties, following exposure to a particular agent, taking into account the inherent characteristics of the agent of concern as well as the characteristics of the specific target system. The Risk Assessment process includes four steps: hazard identification, hazard characterisation (related term: dose-response assessment), exposure assessment, and risk characterisation. It is the first component in a risk analysis process. The definition of risk assessment may vary between Member countries.

Robust(ness): The insensitivity of test results to departures from the specified test conditions when conducted in different laboratories or over a range of conditions under which the test method might normally be used. If a test is not robust, it will be difficult to use in a reproducible manner within and between laboratories.

SAR: See (Q)SARs.

A screen/screening test is often a rapid, simple test method conducted for the purpose of classifying substances into a general category of hazard. The results of a screening test generally are used for preliminary decision making in the context of a testing strategy (*i.e.*, to assess the need for additional and more definitive tests). Screening tests often have a truncated response range in that positive results may be considered adequate to determine if a substance is in the highest category of a hazard classification system without the need for further testing, but are not usually adequate without additional information/tests to make decisions pertaining to lower levels of hazard or safety of the substance

Sensitivity: The proportion of all positive/active substances that are correctly classified by the test. It is a measure of accuracy for a test method that produces categorical results, and is an important consideration in assessing the relevance of a test method.

Specificity: The proportion of all negative/inactive substances that are correctly classified by the test. It is a measure of accuracy for a test method that produces categorical results and is an important consideration in assessing the relevance of a test method.

Standard Operating Procedures (SOP): Formal, written procedures that describe in detail how specific routine, and test-specific, laboratory operations should be performed. They are required by GLP.

Substitute test: A new or revised test that is proposed for use in lieu of a currently used test, regardless of whether the test is being used as a screen, definitive test, or adjunct test, but which may not be able to completely replace the current test method for all testing situations and all substances.

Surrogate: A test species or system used in the place of another target species or test system.

Target species: The species for which information on the potential toxicity or effects of a substance is sought.

Test (or assay): An experimental system used to obtain information on the adverse effects of a substance. Used interchangeably with assay.

Test battery: A series of tests usually performed at the same time or in close sequence. Each test within the battery is designed to complement the other tests and generally to measure a different component of a multi-factorial toxic effect. Also called base set or minimum data set in ecotoxicological testing.

Test method development: The research process before validation in which a test protocol is developed and standardized. The process of test method development should ideally lead to a protocol which is sufficiently detailed and comprehensive to enable the test to undergo prevalidation.

Test method: A process or procedure used to obtain information on the characteristics of a substance or agent. Toxicological test methods generate information regarding the ability of a substance or agent to produce a specified biological effect under specified conditions. Used interchangeably with "test" and "assay".

Test method sponsor: An entity that has proposed a test method for incorporation into a Test Guideline, and for eventual regulatory consideration and acceptance. The test sponsor is usually responsible for organising and providing all the available, relevant information to the organisation overseeing or reviewing the validation status of the test method, or to the agencies which are being asked to accept the test.

Tiered testing approach: Testing which uses test methods in a sequential manner; the test methods selected in each succeeding level are determined by the results in the previous level of testing.

Transferability: The ability of a test procedure to be accurately and reliably performed in independent, competent laboratories.

Validated test method: A test method for which validation studies have been completed to determine the relevance (including accuracy) and reliability for a specific purpose. It is important to note that a validated test method may not have sufficient performance in terms of accuracy and reliability to be found acceptable for the proposed purpose.

Valid test method: A test method considered to have sufficient relevance and reliability for a specific purpose and which is based on scientifically sound principles. A test method is never valid in an absolute sense, but only in relation to a defined purpose.

Validation: The process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose.

Validation management group (VMG): The independent oversight group, consisting of individuals with experience with the types of assays being performed, biostatisticians, and others with knowledge of the purpose of the prevalidation and multi-laboratory phases of the validation study, which is responsible for approving the design and implementation of the study and the selection of the participating laboratories, and co-ordinating the evaluations of the study results. A VMG is also called Management Committee or Management Team.

ANNEX II

EXAMPLES OF VALIDATION STUDIES WITHIN DIFFERENT AREAS OF TEST METHOD DEVELOPMENT

In this Annex, five examples of validation studies within different areas of test method development are presented. These examples are intended to illustrate how different approaches have been used in the validation of test methods that have been proposed and accepted for regulatory use. The examples also provides good reasons why the validation process should be kept flexible and on a case-to-case basis. One goal of this Guidance Document is to present a common understanding of how validation in *in vitro*, *in vivo* and ecotoxicity testing was addressed; and as a second goal, the Guidance Document presents and advocates common validation principles while also recognizing distinctions that can be applied to any area of toxicology, hazard identification/assessment, and risk assessment.

The ecotoxicology examples (1 and 2), for instance, describe a process with an emphasis on standardization of a method rather than on determining the relevance and sensitivity of the method. By including the actual species of interest, ecotoxicity tests incorporate endpoints which are intrinsically relevant and their sensitivity inherent. Therefore, the elements of validation that should be demonstrated are reduced primarily to those of reproducibility. Inter-laboratory (ring) testing functions to verify the reproducibility of the test method and provide an empirical basis for standardizing a Test Guideline and establishing its suitability for regulatory use.

The third example on *in vitro* tests for skin corrosion that have resulted in the acceptance of TGs 430 and 431 gives practical guidance on a variety of approaches to validation. Including the steps of prevalidation and validation, but also a special study on tiered testing approaches in addition to a catch-up validation study.

The *in vivo* Local Lymph Node assay (example 4) represents an alternative test method to TG 406 offering refinement and the validation process is outlined. The fifth example (updated TG 425) is a replacement test for the Acute Oral Toxicity Study (TG 401) where the testing is based on a staircase testing approach.

Example 1:	The OECD *Lemna* growth inhibition test. Development and ring-testing of draft OECD Test Guideline.
Example 2:	Report of the Final Ring Test of the Daphnia magna Reproduction Test.
Example 3:	Validation, Special Study and "catch-up" Validation of *In vitro* Tests for Skin Corrosion Potential – The Human Skin Model Test
Example 4:	The Local Lymph Node Assay (LLNA)
Example 5:	The Up-and-Down Procedure for Acute Oral Toxicity (TG 425)

Example 1

The OECD *Lemna* growth inhibition test: development and ring-testing of a draft OECD Test Guideline.

Environment Agency, Research and Development Technical Report EMA 003, 90 pp., by I. Sims, P. Whitehouse, and R. Lacey, WRc plc (Contractor Report No. EA 4784), Henley Road, Medmenham, Marlow, Buckinghamshire, UK (1999).
Funding provided by Environment Agency, Bristol, UK and USEPA Office of Prevention, Pesticides and Toxic Substances, Washington, DC.

1. Introduction and Rationale for the Proposed Test Method Validated

In a review of Test Guideline (TG) requirements in 1995 (See DRP on Aquatic Testing Methods for Pesticides and Industrial Chemicals No. 11, OECD Environmental Health and Safety Publications, Series on Testing and Assessment, 1998) an aquatic higher plant toxicity test method was highlighted as a high priority for development by the OECD. This led to an initiative by KEMI (Sweden) to develop a new Test Guideline based on aquatic macrophytes of the genus *Lemna*, which was subsequently taken forward by WRc (UK). Inhibition of growth is measured (as changes in yield or biomass and in growth rate) in these floating vascular plants over a period of seven days.

An expert panel or workgroup (steering group) was established in early 1997. Workgroup meetings were held in March and November 1997. At the first meeting research needs that would enable a new OECD TG to be written, were identified and discussed. A “Phase 1", or prevalidation study, was then conducted in 1997 and 1998 by members of the panel to fill in these needs or knowledge gaps. Areas addressed included choice of test species, test media, endpoints, and data analysis. Results of Phase 1 were discussed at the second panel meeting, and work on a draft OECD TG began. Plans for “Phase 2,” an international ring-test or validation study were discussed. A draft test method (June 1998) was then developed and this would form the basis for the method to be validated in Phase 2. The ring-test was conducted in 1998-1999, with financial support provided by the UK Environmental Agency and the US EPA, Office of Prevention, Pesticides and Toxic Substances. A final report was completed later in 1999. A draft of this report had been reviewed by members of the panel. No formal peer review of the ring-test was conducted.

Another draft of this method was developed by the OECD Secretariat in October 2000, with approval by the panel, and distributed to Member countries for comment. Objections were raised by some Member countries on the test endpoints and statistical analysis sections of the draft. Resolution of these objections delayed approval of the proposed new TG. None of these objections had anything to do with aspects of validation or peer review. Discussions and commenting on several more draft TG’s then occurred. After an expert consultation meeting were held in October 2003, agreement among experts from Member countries was achieved and this TG will appear in 2004 as TG 221 in the 16th Addendum of the OECD Test Guidelines Programme.

The TG 221 will be used to assess the effects of pesticides and industrial chemicals on floating aquatic vascular higher plants, and will play an important role in regulation of these substances. It will serve as an important supplement to algal tests, and a *Lemna* test will be a useful addition directly applicable in tiered testing schemes used by Member countries for hazard identification/assessment and risk assessment. Exactly how this TG will be used in regulatory processes will be left to each Member country to decide.

2. Test Method Protocol Components

The principle of the test method is as follows. Exponentially growing plant cultures of the duckweed genus *Lemna* are allowed to grow as monocultures in different concentrations of the test substance over a period of seven days. The objective of the test is to quantify substance-related effects on vegetative growth over this period based on assessments of selected measurement or response variables (frond number, and one other variable for yield such as total frond area, dry weight or fresh weight). To quantify substance-related effects, growth in the test solutions is compared with that of the controls and the concentration bringing about a specified x% inhibition of growth (*e.g.*, 50%) is determined and expressed as the ECx (*e.g.*, EC50). In addition, the lowest observed effect concentration (LOEC) and the no-observed effect concentration (NOEC) may be statistically determined.

The method followed in the ring-test was the draft OECD "*Lemna* growth inhibition test planned ring-test", 1998, Appendix C in the Environment Agency Technical Report cited above. This method will not be discussed in detail here. As a result of the ring-test a revised draft was written (draft OECD Lemna growth inhibition test (JUNE 1999)), which is the Appendix B of the Technical Report. Current, complete details on this method, with subsequent changes made since the 1999 draft, are given in the OECD publication: OECD Guidelines for the Testing of Chemicals – Revised proposal for a new Guideline 221: *Lemna sp.* Growth inhibition test [ENV/JM(2004)33/ANN4], October 22, 2004, and also will not be repeated here.

3. Substances Used for Validation of the Proposed Test Method

After considerable debate among the steering group members, two chemicals were chosen for use in the ring test: potassium dichromate and 3,5-dichlorophenol. These chemicals were selected because:

1) Their effects on the growth of *Lemna* are well-documented.
2) They have different modes of toxic action, with low-moderate toxicity so that use of higher concentrations helps to minimize errors in preparation of solutions and analysis.
3) Stock solutions could be prepared without the use of solvents, and water solubility is "well above" toxic concentrations.
4) They are stable over the test period of seven days, and hydrolytically stable over several months (stock solution).
5) Low volatility from water, and stable/"conservative" speciation over the pH range recommended for the test.
6) Ease of analysis and readily available to all ring-test participants.
7) They represent two well-defined groups of toxicants, *i.e.*, metals and organics, with a "reasonably well-defined" dose-response.

Both chemicals have been widely used in ring-tests and as reference toxicants in tests with other organisms. Since the dose-response appears to be "fairly shallow" the selection of test concentrations was especially important. The test concentrations were chosen to bracket the EC50 with an interval between concentrations not to exceed x3.2.

4. *In Vivo* Reference Data used for an Assessment of the Accuracy of the Proposed Test Method (where appropriate)

A reference test method was not used for comparisons because the *Lemna* test is a new test method, which was based on a composite of several existing methods used by different Member countries. While test data

were available from each of these national test methods there was great uncertainty on which data to use, such that it was found to be impractical to make these comparisons. In the ring-test, results from the controls were compared with the treatments. Performance characteristics assessed included the extent to which validity (doubling time) criteria were met, the intra-laboratory repeatability (used to describe the precision of the method), and the inter-laboratory reproducibility (used to assess the accuracy with which toxicity is estimated). Thirty-seven laboratories from 15 Member countries participated in the ring-test.

5. Test Method Data and Results

The 37 participating laboratories were allowed to test two species of duckweed, *Lemna minor* or *Lemna gibba*. More of the laboratories submitted data for the first species than for the second species, and potassium dichromate was used more extensively in the ring-test than was 3,5-dichlorophenol. Many laboratories repeated tests with the same toxicant and species. Adherence to the 1998 draft TG was good. All participants supplied EC50 data calculated from average specific growth rates as based on frond number, as required by the draft TG, and most (83%; n=191) met the test acceptability criterion for control doubling time. Those tests that failed this criterion were usually conducted under conditions of low illumination or low temperature, or both. Few exceedences of water quality criteria occurred, although the most common problem was an increase in pH in the control medium that was greater than that advised in the 1998 draft. However, this did not seem to be associated with reduced growth rates.

The ranges of measured EC50 values for the two species of *Lemna* and the two test substances are summarized:

Inter-quartile ranges of EC50 values generated in the *Lemna* inter-laboratory ring-test

Test Substance	*Lemna minor* (mg/L)	*Lemna gibba* (mg/L)
Potassium dichromate	2-4 (n=61)	8-30 (n=28)
3,5-dichlorophenol	2.7-3.4 (n=52)	6.00-7.00* (n=4)

* = range of reported EC50 values.

These data were analyzed to identify the contributions to variability in EC50 values resulting from intra-, and inter-laboratory variability. Possible sources of variability were partitioned by using Residual Maximum Likelihood (REML) procedures, with log-transformed EC50 data. Those EC50 values from tests where the doubling time for control frond number was exceeded or where the EC50 was extrapolated beyond the concentration range tested were excluded from the analysis. Both sources of variability were greater in experiments using potassium dichromate but the inter-laboratory variability generally accounted for a greater proportion of the observed variability than the intra-laboratory variability. In tests with *L. minor* and potassium dichromate, two laboratories were responsible for a high proportion of the inter-laboratory variability, suggesting their EC50 estimates were “rather extreme.”

Participating laboratories obtained their own potassium dichromate or 3,5-dichlorophenol, or both, and prepared their own stock solutions for these chemicals.

6. Test Method Performance (Accuracy)

Comparisons with a reference method were not made since in ecotoxicological testing there is generally no accepted reference value. The closest to a reference value in ecotoxicology is where a weight-of-evidence approach is taken. Ecotoxicity test values derived from testing methods with "similar" species may cluster together to give a mean toxicity value. A value from another method may fall within this cluster. Thus, this approach may give some crude indication of method accuracy.

Accuracy has been defined by ICCVAM as determinations of concordance, sensitivity, specificity, positive and negative predictivity, and false positive and negative rates. These various determinations were not made in the *Lemna* ring-testing.

7. Test Method Reliability (Repeatability/Reproducibility)

Quality control criteria were derived and expressed in terms of the deviation from the "consensus" mean EC50 (the geometric mean of the mean EC50s obtained by each laboratory). Estimates of inter-laboratory variance were calculated. When used in conjunction with reference toxicant data, these criteria may be used by laboratories to assess the accuracy with which they perform *Lemna* growth inhibition tests. They may also be used by regulatory authorities when assessing the quality of data to be used in hazard identification/assessment and in risk assessment of chemicals, so that excessive bias and variability can be identified. For EC50 values in tests using *L. minor* and potassium dichromate, the recommended accuracy criteria are 0.6-7.2 mg/L and for this species and 3,5-dichlorophenol, the criteria are 1.7-5.7 mg/L. These ranges serve as "targets" for judging accuracy of *L. minor* tests with potassium dichromate.

Quality control criteria could not be established for *Lemna gibba* because there was a lack of suitable data. Considerably fewer laboratories performed ring-tests with this species (see table above). Because of this lack, very wide control limits would be calculated.

Quality control criteria can also be derived for precision, expressed as the range of EC50 values generated within a laboratory. Estimates of inter-laboratory variability are used, and REML-analysis provided an estimate of the underlying repeatability of duckweed growth inhibition tests for the two toxicants. A series of tests with the reference toxicant should be performed. The variance of the EC50 values from the series of repeat tests is compared with the underlying variance that would be expected, based on the ring-test. This comparison may be done using a chi-square test and an appropriate significance level. Chi-squares were determined for different numbers of repeat tests, based on a given significance level, and multiplier values were calculated. If the variance of EC50 values from a series of tests in a laboratory exceeded the target for intra-laboratory repeatability by more than the calculated multiplier, there would be a high likelihood that this laboratory failed to adequately control variability during testing.

8. Test Method Data Quality

As indicated above, adherence to the 1998 draft test method was judged to be very good. This draft adheres to and is in compliance with OECD GLP guidelines (see OECD 1998 publication: Principles of Good Laboratory Practice and Compliance Monitoring, ENV/MC/CHEM(98)17).

9. Other Pertinent Scientific Reports and Reviews

At this time, it has not been possible to identify any other published or unpublished studies conducted by using the proposed test method.

10. Animal Welfare Considerations (Refinement, Reduction, and Replacement)

These considerations are not relevant to this method and ring-testing.

11. Practical Considerations (Evaluation of strengths and limitations of the test method)

Not relevant because an *in vivo* reference test method was not compared with the proposed new method. See 4.0 above.

12. References

References are listed throughout this document.

13. Supporting Materials (Appendices)

None

Example 2

Report of the Final Ring Test of the *Daphnia magna* Reproduction Test.

OECD, Environment Directorate, OECD Environmental Health and Safety Publications Series on Testing and Assessment No. 6, 190 pp., by I. Sims, P. Van Dijk, J. Gamble, N. Grandy, and M. Huet (1997).

OECD, and Inter-Organization Programme for the Sound Management of Chemicals (IOMC)

1. Introduction and Rationale for the Proposed Test Method Validated

OECD Test Guidelines for Testing of Chemicals are periodically reviewed in relation to scientific progress made since the TG first appeared. Scientific progress is difficult to succinctly define. However, it can include development of new or modified procedures that are accepted by a consensus of ecotoxicologists for incorporation into the test method, an accumulation of new test data and new ways of viewing and using the method, limitations noted with the method as it is used more and more often, and the development of new technology applicable to the test method. For Test Guideline 202, Part II, *Daphnia sp*. Reproduction Test (first adopted in April 1984), it was generally acknowledged that data from tests performed according to this TG could be variable, and that there was a need to further update TG 202.

Results from a European Union (EU) ring-test conducted in 1985, which was based on a revision of the 1984 OECD Test Guideline 202, Part II, also indicated an unacceptable level of variability in the inter-laboratory results. Thirty-seven laboratories in 13 member countries participated. These results led to a series of investigations (*e.g.,* on daphnid genotype, food, and culture medium to use in the method) begun in 1990, initially within the EU and later in OECD Member countries, to identify the sources of variability and then eventually develop an improved Test Guideline. This work was equivalent to a Phase 1 or prevalidation activity.

A series of workshops were held in 1989, 1991, 1993, and 1995 to discuss daphnid toxicity and the reproduction method, how to validate it, and assess the results of validation. OECD became involved after the first workshop.

Results of the investigations were reviewed at the 1991 workshop, and it was determined that further work would be necessary, *i.e.,* a multi-staged ring-test would need to be completed. The first stage would be a pilot or limited ring-test. In 1992 this study (also equivalent to Phase 1 or prevalidation) was conducted with 36 laboratories, from 14 Member countries. Results of this work were reviewed at the 1993 workshop, and planning for a second stage of Phase 2 full ring-test began.

In 1994 a full or final ring-test took place, based on a draft TG 202, Part II (dated February 1994) that was developed after the investigations above were concluded. Forty-eight laboratories in 16 Member countries (and the Czech Republic) participated in this ring-test. Details and results of the ring-test are given in the 1997 Testing and Assessment Report cited above. Based on an analysis of the results of this activity at the 1995 workshop, the test method was again re-drafted. Additional discussion and review occurred in 1996-1997. The test method, as new OECD TG 211, was approved and adopted in September 1998.

Peer review of the pilot ring-test and full ring-test results occurred only among workshop members and attendees.

A daphnid reproduction test is an important higher tier test, which can serve as an indicator or surrogate test species for assessing the effects of pesticides and industrial chemicals on freshwater aquatic invertebrates. This is an important test in hazard identification/assessment, classification and labelling, and in risk assessment.

2. Test Method Protocol Components

The principle of the test is to assess the effect of chemicals on the reproductive output of *Daphnia magna*. Young female daphnids (the parent animals), aged less than 24 hours at the beginning of the test, are exposed to the test substance added to water at a range of concentrations. Test duration is 21 days and at the conclusion of the test, the total number of living offspring produced per parent animal alive at the end of the test is assessed. Juveniles produced by adults that die during the test are excluded from the calculations. Reproductive output of the parent animals can be expressed in other ways (*e.g.*, number of living offspring produced per animal per day from the first day offspring were observed), but these should be reported in addition to the total number of juveniles produced per parent alive at the end of the test. The reproductive output of the animals exposed to the test substance is compared to that of the control(s) to determine the lowest-observed-effect-concentration (LOEC) and the no-observed-effect-concentration (NOEC). In addition, and as far as possible, these data are analyzed by using a regression model to estimate the concentration that would cause an x% reduction in reproductive output, *i.e.*, ECx (*e.g.*, EC50, EC20, EC10). Also, measured are survival of the parent animals and time to production of first brood. Other substance-related effects on parameters such as growth (*e.g.*, length), and possibly intrinsic rate of increase, may also be examined.

The method followed in the full ring-test was the draft OECD Test Guideline 202, part II, *Daphna magna* reproduction test to be used in the final ring test, Appendix B in the 1997 OECD report cited above. This method will not be discussed in detail here. Current complete details on this method, with subsequent changes made since the Appendix B draft TG, are in the approved OECD TG 211 *Daphnia magna* Reproduction Test, adopted on September 21, 1998.

3. Substances Used for Validation of the Proposed Test Method

Three test substances were used in the final ring-test, 3,4-dichloroaniline (DCA), cadmium chloride, and phenol. These chemicals were chosen because:

1) They have well-documented effects on *Daphnia* reproduction;
2) They have different modes of toxic action;
3) Stock solutions could be prepared without using solvents;
4) These chemicals would be relatively easy and inexpensive to analyze at the concentrations to be used.

DCA was chosen to provide a link with the 1985 EU ring-test and the 1992 pilot ring-test. Cadmium was chosen to investigate the suitability of two culture media, which contain a known chelating agent, for testing metals and metal compounds. Phenol was selected to assess the performance of the draft TG when testing a “difficult” substance because it is readily biodegradable in the test system.

4. *In Vivo* Reference Data used for an Assessment of the Accuracy of the Proposed Test Method (where appropriate)

In vivo reference data and the method used to generate these data were not used to assess the accuracy of the February 1994 method used in the full ring-test. Also, the February draft of TG No. 202, Part II, differed substantially from the 1984 version, and it would thus have been difficult to compare one with the

other. However, the results of this ring-test were compared with those from the 1985 ring-test, when the guideline was less defined, and were also used to investigate the intra or inter-laboratory variability.

5. Test Method Data and Results

The final ring-test thoroughly evaluated the performance of the February 1994 draft of TG 202, Part II. Variability in the test results within and among laboratories was "low." Significant improvements to the TG have been made since 1984 as a result of this testing. In addition, it was possible to investigate various ways of expressing reproductive output of the daphnids, and how to treat data from animals that die during the test. Proposals for how these issues should be dealt with were made. Adherence to the instructions in the February 1994 draft by the laboratories conducting the full ring-test was generally "very good." "Most" laboratories were able to meet the recommended water quality criteria. Of the 98 experiments (52 with DCA, 30 with cadmium, and 16 with phenol), the dissolved oxygen concentration, a critical test condition, fell below the minimum in only one test. Also, 85% of the experiments were conducted within the specified temperature range.

For DCA, approximately 50% of the NOEC and ECx values differed by a factor of 2 or less, over 75% differed by a factor of 4 or less, and over 90% differed by a factor of 8 or less. For cadmium, 38% of the NOECs differed by a factor of 2 or less and 62% differed by a factor of 8 or less. Cadmium EC50 s differed by 27% and 45% respectively. For phenol, 45% of the NOECs differed by a factor of 3.2 or less, and 82% differed by a factor of 10 or less. Phenol EC50 s differed by 70% and 100%, respectively.

"Most" laboratories met the validity control performance criteria in the draft TG: 98% of the experiments reported control mortalities of not greater than 20%, and 84% of experiments reported mean numbers of live offspring per parent of at least 60.

6. Test Method Accuracy (Performance)

Comparisons with a reference test method were not made since in ecotoxicological testing there is no generally accepted reference value. The closest to a reference value in ecotoxicology is where a weight-of-evidence approach is taken. Ecotoxicity test values derived from testing methods with "similar" species may cluster together to give a mean toxicity value. A value from another method may fall within this cluster. Thus, this approach may give some crude indication of method accuracy.

Accuracy has been defined by ICCVAM as determinations of concordance, sensitivity, specificity, positive and negative predictivity, and false positive and negative rates. These various determinations were not made in the Daphnia magna ring-testing.

7. Test Method Reliability (Repeatability/Reproducibility)

An important objective of the full ring-test was to obtain information on intra-laboratory variability, and to assess how this compares with inter-laboratory variability. Laboratories were asked to repeat the tests under the same test conditions, where possible. Statistical analyses were performed to estimate intra-, and inter-laboratory variability for EC50 results using nominal concentrations of DCA only (the other two chemicals had insufficient data sets for this analysis). The analysis was limited to EC50 values using total juveniles as the response variable or test endpoint, for only laboratories that had performed repeat tests. Results from ten laboratories were compared by using these criteria.

Intra or inter-laboratory variability (repeatability) was defined as the value below which the absolute difference between two single test results obtained at the same laboratory, under repeatable conditions,

may be expected to lie with a probability of 95%. Intra-laboratory variance was estimated to be 12.36, and using square roots to convert variances to SD gave a value of 3.5.

Inter-laboratory variability (reproducibility) was defined as the value below which the absolute difference between two single test results obtained at different laboratories, under reproducible conditions, may be expected to lie with a probability of 95%. Inter-laboratory variance was calculated to be 38.40, and using square roots to convert variances to SD, gave a value of 6.2.

The repeatability and reproducibility values were calculated by multiplying the 3.5 or 6.2 values by the square root of 2 to give the standard error of the difference between two observations and then by 1.96 (the 95% two sided point of the standard normal distribution). The combined multiplier was calculated to be 2.8. Using this multiplier yielded a repeatability value of 10µg/L and a reproducibility value of 20 µg/L, with a ratio of 2. Thus, repeat results among laboratories are twice as variable as repeat results within a laboratory. Furthermore, two EC50 results from within the same laboratory cannot be considered to be different unless they differ by more than 10 µg/L, and two results from different laboratories cannot be considered different unless they differ by more than 20 µg/L. This ratio of 2 compares favourably with ratios obtained from other ring tests. A fathead minnow ring-test had a ratio of 1.3, and a Daphnia 48 hour acute toxicity ring-test had a ratio of 2.6.

8. Test Method Data Quality

As indicated above, adherence to the February 1994 draft method was generally "very good." This draft adheres to and is in compliance with OECD GLP guidelines (see OECD 1998 publication: Principles of Good Laboratory Practice and Compliance Monitoring, ENV/MC/CHEM(98)17).

9. Other Pertinent Scientific Reports and Reviews

At this time, it has not been possible to identify any other published or unpublished studies conducted by using the updated test method.

10. Animal Welfare Considerations (Refinement, Reduction, and Replacement)

These considerations are not relevant to this method and ring-testing.

11. Practical Considerations (Evaluation of strengths and limitations of the test method)

Not relevant because an *in vivo* reference test method was not compared with the updated new method. See 4.0 above.

12. References

References are listed throughout this document.

13. Supporting Materials

None

Example 3

Validation, Special Study and "catch-up" Validation of *In vitro* Tests for Skin Corrosion Potential – The Human Skin Model Test

Summary: *After a series of studies, in 1998, two in vitro tests for skin corrosivity, namely the rat skin transcutaneous electrical resistance (TER) assay and a test employing a human 3D reconstructed skin model, had been endorsed as scientifically valid by the European Centre for the Validation of Alternative Methods (ECVAM, Joint Research Centre, European Commission). As a consequence of requirements specified in EU Directive 86/609/EEC for the protection of laboratory animals, the two methods were in the year 2000 adopted as one new Test Method B.40 ("In vitro Skin Corrosion") of Annex V to EU Directive 67/548/EEC (on dangerous substances). However, at the more global OECD level the two methods had not reached consensus, mainly because descriptions were too specific, requiring exactly the experimental set-up's used in the validation studies. By the end of 2001, two OECD Extended Nominated Expert Consultations revealed that both methods could be described more general allowing for more flexibility if additional conditions / requirements specified in the newly proposed Test Guidelines TG 430 (TER) and TG 431 (human skin model) were met. Briefly, prevalidation-, validation, and additional special studies were performed on a total of five in vitro tests for skin corrosion. In the prevalidation study three tests, the rat skin TER test, the CORROSITEX test, and the Skin² human skin model test were evaluated in 1994-1995. Fifty coded chemicals were tested in two to three laboratories. All tests showed sufficient reproducibility and a promising concordance with in vivo skin corrosion potential. Experts reviewing the outcome in ECVAM Workshop No 6 recommended proceeding to a formal validation study after method refinements were made. The selection and acquisition of 60 new reference chemicals backed high quality in vivo rabbit test data took almost a year. The formal validation study was then conducted with the three tests of the prevalidation study, plus a second skin model test (EPISKIN) in 1996-1997, and evaluated and published 1998. Whereas CORROSITEX showed a limited applicability and an insufficient predictivity, and the Skin2 test protocol used (accepted by US DOT) was too insensitive, the TER test and the EPISKIN test both met the predefined acceptance criteria. During the validation study, it was decided to perform an additional small-scale special study: Since for the 60 test chemicals SAR data, in vivo data and in vitro data (from the validation study) were available, it was decided to additionally determine pH and acid-alkaline reserve, and apply the OECD tiered testing approach outlined in OECD TG 404. This study showed that by SAR, pH and in vitro data 100% of the corrosive chemicals were detected. Since both Skin2 and EPISKIN became unavailable due to marketing decisions of the producers in 1996-1997, ECVAM supported a new type of "catch-up" validation study. This catch-up study was designed to determine if another human skin model (EpiDerm; available since 1992) could be used with a test protocol resembling the one used for the successfully validated EPISKIN. The outcome of the study was almost identical with EPISKIN, confirming the robustness of the test design and the assumption that OECD Test Guideline 431 can be applied to different (including new) skin models.*

1. Introduction and Rationale for the Proposed Test Method Validated

Skin corrosion refers to the production of irreversible tissue damage in the skin following the application of a test material as defined by the United Nations Globally Harmonised System for the Classification and Labelling of Chemical Substances and Mixtures (GHS). The assessment of skin corrosivity has typically involved the use of laboratory animals (rabbits). Concern for the pain and suffering involved with this procedure has been addressed in the 2002 revision of OECD Test Guideline 404 allowing for the determination of skin corrosion by using alternative *in vitro* methods.

A first step towards defining alternative tests that could be used for skin corrosivity testing for regulatory purposes was the conduct of prevalidation studies (1). Following this, a formal, validation study of *in vitro* methods for assessing skin corrosion (2)(3) was conducted (4)(5)(6). The outcome of these studies and

other published literature (7) led to the recommendation that two *in vitro* tests could be used for the assessment of *in vivo* skin corrosivity (8)(9)(10)(11): the human skin model test (this Method) and the transcutaneous electrical resistance test (OECD Test Guideline 430). The validation studies have reported that tests employing human skin models (5)(6)(7) are able to reliably discriminate between known skin corrosives and non-corrosives, in particular, if supported by physicochemical information (12).

2. Test Method Protocol Components

The mechanisms of skin corrosion *in vivo* are well understood. One mechanism is a destructive "digestion" of the skin barrier and underlying viable skin cells by pH extremes (acids and alkalines and their derivates), the other main mechanism is the combination of high cytotoxicity / reactivity of a substance with high skin permeability. In this case, the toxic substance quickly reaches the basal skin layers revealing a necrotic (irreversible) damage. Both *in vivo* mechanisms can be detected in a specific short term protocol of a cytotoxicity test, measuring cellular viability in a human skin model (provided it has a well developed stratum corneum barrier).

The test material is applied topically to a three-dimensional human skin model, comprising at least a reconstructed epidermis with a functional stratum corneum. Corrosive materials are identified by their ability to produce a decrease in cell viability (as determined, for example, by using the MTT reduction assay (13)) below defined threshold levels at specified exposure periods.

3. Substances Used for Validation of the Proposed Test Method

Prevalidation Study (1993-1994): Fifty test chemicals were selected on the basis of their availability and the confidence with which they could be classified unambiguously as "corrosive"or "non-corrosive" (1). Most of the test chemicals (25 corrosives, 25 non-corrosives) were commercially available.All test samples were independently coded by Unilever ESL and codes were unknown to the participants. To ensure the integrity of the raw data, test results were lodged with ECVAM prior to the code being broken.

Formal Validation Study (1996-1998): Sixty chemicals (30 corrosives / 30 non-corrosives) were selected representing all chemical classes relevant to skin corrosion potential: organic acids, organic bases, neutral organics, phenols, inorganic acids, inorganic bases, inorganic salts, electrophiles, and soaps/surfactants (2). The main sources for the selection were *in vivo* data from ECETOC Report No. 66, but also *in vivo* data sponsored from In Vitro International (IVI, Irvine, USA). All test samples were independently coded and distributed by BIBRA Intl. and codes were unknown to the participants. To ensure the integrity of the raw data, test results were lodged with ECVAM prior to the code being broken.

Special Study on OECD TG 404 Tiered Testing Approach (1997-1998): In the "special study" the same set of 60 chemicals as used in the formal validation study was evaluated for pH, and acid/alkaline reserve in a blinded manner (12).

Catch-up Validation Study (1997-1999): Fifty chemicals (25 corrosives / 25 non-corrosives) were used to adapt and refine the EpiDerm protocol and Prediction Model, the main selection criterion being the lowest possible overlap of substances with the formal ECVAM validation study. In the final blind trial (performance assessment in three laboratories), a representative sub-set of the formal validation study was used (7).

4. *In vivo* Reference Data used for an Assessment of the Accuracy of the Proposed Test Method (where appropriate)

Prevalidation Study (1993-1994): In vivo data were obtained from in- house studies or from the scientific literature, or by reference to the 1991 "Comité Européen des agents de Surface et leurs Intermediaries Organiques" (CESIO) classifications, or to manufacturers data. All animal tests were reported to have been conducted in accordance with OECD Test Guideline 404. Corrosivity classifications were from three sources: animal data, hazard data sheets, or Annex 1 of the Dangerous Substances Directive (67/548/EEC) The selection of the test chemicals was based primarily on the availability of the test material and associated data. Thus, the criteria for selection were not as stringent as those employed by other groups (for example, the European Centre for the Ecotoxicology and Toxicology of Chemicals [ECETOC].

Formal Validation Study (1996-1998): An independent "Chemical Selection Sub-Committee" (CSSC) was set up to guarantee that chemicals selected met the highest standards with regard to their *in vivo* classification, supporting further evidence on corrosive/non-corrosive properties, chemical classes represented, corrosivity subcategories represented, and avoidance of a majority of pH-extremes that could be identified as corrosives by other means than a biological test system. In the CSSC the European chemical industry (ECETOC) was represented, as well as a competent authority, an SAR expert, and the sponsor, ECVAM. Sixty chemicals (30 corrosives/30 non-corrosives) were selected representing all relevant chemical classes: organic acids, organic bases, neutral organics, phenols, inorganic acids, inorganic bases, inorganic salts, electrophiles, and soaps/surfactants. A chemical was only selected if at least one original *in vivo* study report was available. The main source was data from ECETOC Report No. 66.

Reference Data proposed in OECD Test Guideline 431:

To allow future new "me too" skin models to be used in the context of the accepted OECD Test Guideline 431, twelve chemicals were selected by Experts in an OECD Extended Nominated Expert Consultation (11). Most of the chemicals listed are taken from the list of chemicals selected for the ECVAM international validation study (3). Their selection is based on the following criteria:

i) equal number of corrosive and non-corrosive substances;

ii) commercially available substances covering most of the relevant chemical classes;

iii) inclusion of severely corrosive as well as less corrosive substances in order to enable discrimination based on corrosive potency;

iv) choice of chemicals that can be handled in a laboratory without posing other serious hazards than corrosivity.

Reference Chemicals list:

1,2-Diaminopropane	CAS-No. 78-90-0	Severely Corrosive
Acrylic Acid	CAS-No. 79-10-7	Severely Corrosive
2-tert. Butylphenol	CAS-No. 88-18-6	Corrosive
Potassuim hydroxide (10%)	CAS-No. 1310-58-3	Corrosive
Sulfuric acid (10%)	CAS-No. 7664-93-9	Corrosive
Octanoic acid (caprylic acid)	CAS-No. 124-07-02	Corrosive
4-Amino-1,2,4-triazole	CAS-No. 584-13-4	Not corrosive
Eugenol	CAS-No. 97-53-0	Not corrosive
Phenethyl bromide	CAS-No. 103-63-9	Not corrosive
Tetrachloroethylene	CAS-No. 127-18-4	Not corrosive
Isostearic acid	CAS-No. 30399-84-9	Not corrosive
4-(Methylthio)-benzaldehyde	CAS-No. 3446-89-7	Not corrosive

5. Test Method Data and Results

For the data obtained in the prevalidation study (1) with the full thickness human skin model Skin² a prediction model developed for classification into 3 corrosivity subcategories (I/II/III) according to UN regulations for the transport of dangerous goods were applied, because this *in vitro* test had been accepted in 1994 by the US DOT based on in house data submitted by the producer of the skin model. However, in the prevalidation study the ability to correctly subcategorise corrosives could not be confirmed, and results were therefore evaluated on the basis of ability to correctly separate the groups of skin corrosives from non-corrosives (C/NC).

In the validation study (3), the two evaluated human skin model tests (Skin² and EPISKIN) used different protocols and different prediction models. While the EPISKIN protocol and prediction model were closely matching the design of the *in vivo* rabbit test (MTT reduction data after exposure of 3 min, 1 hr, and 4 hrs were used to classify C/NC, or R34/R35/NC), for the Skin² test the US DOT accepted protocol and prediction model was used. Expectedly, due to the short exposure of only 10 seconds, the Skin² test was 100% specific, but showed an unacceptable low sensitivity (43%) and did therefore not meet the predefined acceptance criteria for a validated skin corrosion test (3). Information given in the following paragraphs on the outcome of this study is therefore only related to EPISKIN, which showed an acceptable performance.

In the catch-up validation study, performed with the skin model EpiDerm (7), the original prediction model to interpret data derived from a MTT reduction time course assay (14) "*a chemical reducing the MTT conversion after an exposure of less than 3 minutes is a corrosive*" was turned into a prediction model which was closer related to the validated EPISKIN test ("*if the MTT conversion after 3 minutes exposure is below 50% of untreated controls, the chemical is predicted corrosive"*. In a second phase of this study (7), an additional 1 hr exposure was added, to obtain an increased test sensitivity, and a balanced positive and negative prediction of > 80%. A third exposure time of 4 hrs (as used in EPISKIN) was not used, since it did not contribute to an increase of test performance (7).

6. Test Method Accuracy (Performance)

Compared to *in vivo* rabbit data, the following Cooper statistics (2x2 tables) performance measures were obtained in the three studies (1)(3)(7):

Prevalidation study (Skin²):
Sensitivity: 64%, 96%, 84% (laboratory A, B, and C)
Specificity 76%, 88%, 76% (laboratory A, B, and C)
The performance in laboratory B was regarded acceptable, and in laboratories A and C promising to proceed to a formal validation study.

Validation Study (EPISKIN):
Sensitivity: 81% / 83% (overall performance, prediction model A / B)
Specificity 80% / 80% (overall performance, prediction model A / B)
The performance of both prediction models was regarded acceptable (with a slight preference to prediction model B) as the measures exceeded the predefined acceptance criteria of the best possible outcome considering the variability of the *in vivo* reference data.

Catch-up Validation Study (EpiDerm):
Sensitivity: 72% / 88% (overall performance, prediction model A / B)
Specificity 100% / 86% (overall performance, prediction model A / B)
The performance of prediction model A (using only data of 3 min exposure) was regarded not sufficient as for a replacement test (however, valuable for positive screening). The performance of prediction model B (using data of 3 min and 1 hr exposure) was regarded excellent as the measures exceeded the predefined acceptance criteria of the best possible outcome considering the variability of the *in vivo* reference data.

7. Test Method Reliability (Repeatability/Reproducibility)

To check repeatibility over time within one laboratory and reproducibilty between laboratories, in all three studies (1)(3)(7) ANOVA was applied. In case of the formal validation study (3) the reproducibility was even assessed for each of the 60 test chemicals. Repeatability within a laboratory over time was excellent in all three studies and no significant differences were observed between independent tests. With regard to reproducibility, the prevalidation study revealed a higher level of concordance (80%) between laboratories B and C than between A and B, or A and C, indicating that the test protocol needs to be better defined and laboratory trainings improved. In the two following studies, concordance between laboratories was almost 100% and ANOVA revealed that difference between laboratories had no impact on differentiating the test chemicals by their corrosive potential.

8. Test Method Data Quality

Whereas the prevalidation study was performed applying GLP standards, the formal validation study and the "catch -up-Validation Study were performed fully GLP compliant. However, none of the studies was subjected to an internal or external study audits.

9. Other Pertinent Scientific Reports and Reviews

Following the publication of the three studies (1)(3)(7), and the acceptance of the two skin model tests by ECVAM (8)(9), ICCVAM evaluated the methods and published in 2002 a report on the EpiDerm, EPISKIN and TER skin corrosion tests (15)

In 2004, ICCVAM published recommended Performance Standards to be met by new "me too" tests of similar nature as CORROSITEX, TER, EpiDerm, and EPISKIN (16). The requirements to be met by new test systems defined by ICCVAM are higher (with regard to the reference chemical number) than defined in OECD Test Guideline 431.

10. Animal Welfare Considerations (Refinement, Reduction, and Replacement)

For the rare purpose of only assessing skin corrosivity potential, the test may be regarded as Replacement test. However, in almost all cases knowledge of skin irritation potential will be necessary as well. As long as no validated skin irritation test will be available, the test will function as part of the tiered strategy shown in OECD TG 404, to screen out corrosive chemicals, so to act as Reduction and Refinement test.

11. Practical Considerations (Evaluation of strengths and limitations of the test method)

Chemical action by the test material on the vital dye may mimic that of cellular metabolism leading to a false estimate of viability. This has been shown to happen when such a test material is not completely removed from the skin by rinsing (7). If the test material directly acts on the vital dye, additional controls should be used to detect and correct for test substance interference with the viability measurement.

12. References

(1) Botham, P.A., Chamberlain, M., Barratt, M.D., Curren, R.D., Esdaile, D.J., Gardner, J.R., Gordon, V.C., Hildebrand, B., Lewis, R.W., Liebsch, M., Logemann, P., Osborne, R., Ponec, M., Regnier, J.F., Steiling, W., Walker, A.P., and Balls, M. (1995). A prevalidation study on *in vitro* skin corrosivity testing. The report and recommendations of ECVAM Workshop 6. ATLA 23, 219-255.

(2) Barratt, M.D., Brantom, P.G., Fentem, J.H., Gerner, I., Walker, A.P., and Worth, A.P. (1998). The ECVAM international validation study on *in vitro* tests for skin corrosivity. 1. Selection and distribution of the test chemicals. Toxic. in Vitro 12, 471-482.

(3) Fentem, J.H., Archer, G.E.B., Balls, M., Botham, P.A., Curren, R.D., Earl, L.K., Esdaile, D.J., Holzhutter, H.-G., and Liebsch, M. (1998). The ECVAM international validation study on *in vitro* tests for skin corrosivity. 2. Results and evaluation by the Management Team. Toxic. in Vitro 12, 483-524.

(4) OECD (1996). Final Report of the OECD Workshop (Solna Report) on Harmonization of Validation and Acceptance Criteria for Alternative Toxicological Test Methods, 62pp.

(5) Balls, M., Blaauboer, B.J., Fentem. J.H., Bruner. L., Combes, R.D., Ekwall, B., Fielder. R.J., Guillouzo, A., Lewis, R.W., Lovell, D.P., Reinhardt, C.A., Repetto, G., Sladowski. D., Spielmann, H., and Zucco, F. (1995). Practical aspects of the validation of toxicity test procedures. The report and recommendations of ECVAM workshops. ATLA 23, 129-147.

(6) ICCVAM (1997). Validation and Regulatory Acceptance of Toxicological Test Methods. A Report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods, 105pp.

(7) Liebsch, M., Traue, D., Barrabas, C., Spielmann, H., Uphill, P., Wilkins, S., McPherson, J.P., Wiemann, C., Kaufmann, T., Remmele, M. and Holzhütter, H.-G. (2000). The ECVAM prevalidation study on the use of EpiDerm for skin corrosivity testing. ATLA 28, pp. 371-401.

(8) ECVAM (1998). ECVAM News & Views. ATLA 26, 275-280.

(9) ECVAM (2000). ECVAM News & Views. ATLA 28, 365-67

(10) ICCVAM. (2001). Dermal corrosivity assays: EpiDerm™, EPISKIN™ and Rat Skin Transcutaneous Electrical Resistance. Background review document; public comments. Research Triangle Park, NC: Interagency Coordinating Committee on the Validation of Alternative Methods/National Toxicology Program Center for the Evaluation of Alternative Toxicological Methods. http://iccvam.niehs.nih.gov/methods/epiddocs/epis_brd.pdf

(11) OECD (2002) Extended Expert Consultation Meeting on The In Vitro Skin Corrosion Test Guideline Proposal, Berlin, 1st –2nd November 2001, Secretariat's Final Summary Report, 27th March 2002, OECD ENV/EHS, available upon request from the Secretariat

(12) Worth AP, Fentem JH, Balls M, Botham PA, Curren RD, Earl LK, Esdaile DJ, Liebsch M (1998). An Evaluation of the Proposed OECD Testing Strategy for Skin Corrosion. ATLA 26: 709-720.

(13) Mosmann, T. (1983). Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. J. Immunol. Meth. 65, 55-63.

(14) Perkins, M.A., Osborne, R. & Johnson, G.R. (1996). Development of an *in vitro* method for skin corrosion testing. Fundamental and Applied Toxicology 31. 9 - 18

(15) ICCVAM (2002) ICCVAM Evaluation of EPISKIN™, EpiDerm™ (EPI-200), and the Rat Skin Transcutaneous Electrical Resistance (TER) Assay: *In Vitro* Test Methods for Assessing Dermal Corrosivity Potential of Chemicals (NIH Publication No: 02-4502) http://iccvam.niehs.nih.gov/methods/epiddocs/cwgfinal/cwgfinal.pdf

(16) ICCVAM (2004) Recommended Performance Standards for *In Vitro* Test Methods for Skin Corrosion (NIH Publication No: 04-510) http://iccvam.niehs.nih.gov/methods/ps/ps044510.pdf

13. Supporting Materials

None

Example 4

The Local Lymph Node Assay (LLNA)

1. Background

In 1997, Drs. Frank Gerberick (Procter and Gamble, USA), David Basketter (Unilever, UK), and Ian Kimber (Zeneca, UK) proposed to ICCVAM that the murine local lymph node assay (LLNA) should be considered for regulatory acceptance as a substitute for existing test methods used to assess dermal hypersensitivity. The LLNA had undergone validation studies to characterize its utility and limitations (Basketter and Scholes, 1992; Basketter et al., 1991, 1994, 1996; Chamberlain and Basketter, 1996; Gerberick et al., 1992; Kimber and Weisenberger, 1989; Kimber et al., 1989, 1990, 1991, 1995, 1998; Loveless et al., 1996; Scholes et al., 1992). The sponsors asked ICCVAM to evaluate the validity of the LLNA as a stand-alone substitute to guinea pig (GP) methods traditionally used to identify contact allergens. The LLNA was already accepted as a screening test whereby a positive response would be accepted, but a negative response would require GP testing. Their proposed use of the LLNA was that both negative and positive LLNA data should be accepted without the need for secondary testing. Thus, chemicals that fail to elicit a positive response in the LLNA should be classified as lacking significant skin sensitizing potential and should be handled and labelled accordingly without the need for further confirmation testing. The sponsors provide information supporting the validity of the LLNA as outlined in the ICCVAM test method submission guidelines available at that time (ICCVAM, 1999).

2. Scientific Rationale and Mechanistic Basis of the LLNA

Allergic Contact Dermatitis (ACD) develops in two phases. In the first phase, sensitisation (induction phase) occurs after initial skin contact with a sensitising agent. At the site of chemical contact, there is an immediate release of signalling factors and activation of the skin dendritic cells. The dendritic cells process the chemical, and subsequently mature and migrate to the draining lymph node. It is this process that induces the dendritic cells to become effective antigen presenting cells. Lymphocytes within the nodes, upon antigen presentation, undergo cellular proliferation. Proliferation is the endpoint evaluated in the LLNA and the mechanistic basis of the assay. It also marks the first phase of ACD, the induction phase (Selgrade et al., 2001).

Following proliferation, T lymphocytes are considered “primed” as they have a specific recall for the sensitizing agent. Upon additional exposure to the agent, an antigen specific response occurs which is referred to as the elicitation phase. This second phase occurs only upon challenge with the specific agent. There is an immediate release of inflammatory agents, in addition to immune specific mediators that cause an inflammatory cell influx to the dermal site. Elicitation is a systemic response that can occur at locations other than the original site of sensitization. The elicitation phase is characterised by erythema and oedema and occurs 24 to 72 hours after the challenge exposure. This response is the endpoint assessed in traditional GP tests (Selgrade et al., 2001).

3. Regulatory Rationale for the LLNA

ACD is associated with chemical exposure in the workplace and at home. An assessment of the potential for chemicals to cause ACD is an important component of safety testing. Traditionally, GP tests have been used for more than 65 years to assess the ACD potential of chemicals, pharmaceuticals, and consumer products (Landsteiner and Jacobs, 1935; ICCVAM, 1999). While there are variations in these test methods, the guinea-pig maximization test (GPMT) and the Beuhler Assay (BA) are used most commonly. Both of these tests rely on the sensitization and challenge phases of ACD, utilize 20 animals per test group, and require about one month to perform the assay. The GPMT has been criticized for the deleterious effect that

the use of adjuvant has on treated animals; however, this test method is considered more "sensitive" than the BA. The endpoint measured for both GP methods is a visual assessment (subjective) of erythema at the challenge location and requires substantial technical expertise (Marzulli and Maibach, 1996).

4. Description of the LLNA Test Method Protocol

The LLNA measures lymphocyte proliferation in the draining lymph nodes of mice topically exposed to the test article using the incorporation of radioactive thymidine or iododeoxyuridine into DNA. The results are expressed as a ratio (the stimulation index or SI) of the mean number of disintegrations per minute (DPM) for treated mice as compared to controls. Chemicals with a SI of 3.0 or more are considered positive, and those with a SI less than 3.0 are considered negative. This scoring differs from the scoring in GP assays in which a test substance is classified as positive based on the percentage of animals in a group that are responders (at least 15% in a non-adjuvant assay and at least 8% in an adjuvant test) (Marzulli and Maibach, 1996). The measured proliferation in the LLNA is an essential biological response during the induction phase of sensitization, and results from the increased proliferation of lymphocytes in the lymph node due to the sensitizing chemical exposure. In contrast, as mentioned earlier, the currently accepted GP assays, primarily the GPMT and the BA, rely on the elicitation phase of ACD by measuring skin reactivity to a secondary challenge with the test substance. Furthermore, as mentioned earlier, endpoint interpretation in GP tests is subjective and requires substantial technical expertise.

The proposed detailed complete protocol necessary to perform the LLNA was provided. The basis for protocol modifications made during development and validation was described, as was the rationale for critical aspects of the protocol. Examples of other protocol elements described included: the basis for decision criteria to classify a response as a positive or negative result; rationale for selection of the test system (*i.e.* strain, sex, age, weight); rationale for the selection of positive controls; and basis for the dose selection procedures.

LLNA and Reference Data: LLNA data were submitted for 209 chemicals representing a wide range of chemical and product classes. The accuracy of the LLNA for predicting the outcome from standard ACD tests was assessed by comparing LLNA results with reference data for 126 chemicals from standard GP test methods. The accuracy of the LLNA for predicting human ACD was assessed by comparing LLNA results with data from 39 chemicals tested in humans using the human maximization test, and 35 additional chemicals included in human patch test allergen kits. The extent to which data were collected in accordance with Good Laboratory Practice guidelines was described.

Test Method Performance: Calculations of performance were provided for sensitivity, specificity, accuracy, positive and negative predictivity, and false positive and negative rates. The overall accuracy of the LLNA compared to the guinea pig results was 86%. The accuracies of the LLNA and the guinea pig tests, compared to human results, were both 72%.

Intra- and Inter-Laboratory Reproducibility: Data and analyses were provided to characterize the intra- and inter-laboratory reproducibility of the LLNA. Three studies were provided, with each study involving up to six different chemicals in three to six laboratories.

Animal Welfare Considerations: The extent that the LLNA reduced, refined, or replaced animal use compared to the existing GP method was described. The basis for determining that the number of animals used was the minimum number necessary to accomplish the testing objective was provided. The submission stated that the LLNA required fewer animals than the traditional GP tests. It also stated that the LLNA did not involve the pain and distress that might occur with procedures in the GP tests.

5. ICCVAM Evaluation and Peer Review of THE LLNA

ICCVAM determined that the LLNA submission was sufficiently complete and that it was relevant to the needs of agencies and agreed to proceed with an independent peer review panel evaluation of the LLNA. An ICCVAM Immunotoxicity Working Group (IWG) composed of Federal government scientists was formed to assist with the peer review evaluation and implementation process. Working group representatives had specific knowledge and experience applicable to the proposed test method, or were familiar with the current test methods used to generate data for ACD. NICEATM coordinated communications, meeting preparation, document handling, literature reviews, statistical support, and final report preparation. Public comment was encouraged throughout the process. A public docket was established where all LLNA data and information considered by the Panel could be accessed. The public Peer Review Meeting, availability of the test method submissions, and a request for public comment on the LLNA were announced in a Federal Register Notice (63 FR 37405-37406).

6. LLNA Independent Peer Review Panel

The LLNA Panel consisted of 13 national and international experts in the field of immunotoxicology and allergic contact dermatitis (Dean et al., 2001). Panel members were from academia, industry, and government, and had experience in biostatistics, toxicology, clinical medicine, and molecular immunology. The international community was represented with Panel members from Denmark, Japan, and Norway. All Panel members were required to sign a statement indicating that their involvement in the evaluation of the LLNA did not constitute a financial or other conflict of interest. Panel members reviewed extensive submission materials on the LLNA, participated in the public Peer Review Meeting, and prepared a written report of their evaluation of the LLNA.

7. Panel Evaluation of the LLNA

The Panel was charged with evaluating the extent to which the ICCVAM validation and acceptance criteria had been addressed. Questions were developed to ensure the completeness of the Panel's evaluation with regard to test method description, test method performance, test method quality, determination of test method reliability (repeatability/reproducibility), other scientific reviews, other considerations, and related issues. Two overall questions were also proposed to the Panel: (1) Has the LLNA been evaluated sufficiently and is its performance satisfactory to support its adoption as a stand-along alternative? and (2) Does the LLNA offer advantages with respect to animal welfare considerations (refinement, reduction, and replacement)?

The Panel requested additional information and analyses that they considered necessary for their evaluation. Examples of information requested by the Panel included original data, additional statistical analyses, and additional information about the traditional GP assays. LLNA data and comments received from public sources were provided to the Panel for their consideration.

8. Panel Meeting

The Panel convened on September 17, 1998, in Gaithersburg, Maryland, to review the LLNA in a public meeting. Presentations were made on the ICCVAM criteria and process, the current regulatory testing requirements for ACD, and the procedure for conducting the LLNA. Panel members then presented and discussed their evaluations, and presented their conclusions and recommendations. Public comment was permitted at specific times during the meeting. The sponsors provided useful comments and responded to questions from the Panel.

9. Panel Report

The report of the LLNA Panel was published in 1999 (ICCVAM, 1999). Details on the LLNA Panel's conclusions and recommendations can be found also in Dean et al. (2001). The Panel unanimously recommended the LLNA as a stand-alone alternative to GP tests, provided that some specific recommendations on procedural aspects were incorporated (ICCVAM, 1999; Dean et al., 2001). In addition, the Panel concluded unanimously that the LLNA provided animal welfare improvements compared to currently accepted GP methods with respect to animal use, reduction, and refinement.

10. ICCVAM Test Method Recommendations

Following completion of the LLNA Peer Panel Report, the IWG developed and forwarded draft recommendations on use of the LLNA to ICCVAM. The IWG concluded that the LLNA provided specific advantages and potentially more accurate assessments of ACD compared to GP tests. One advantage is that the interpretation of the LLNA is not affected by coloured agents, as may occur with GP tests. Therefore, the LLNA was recommended as the preferred test method for coloured agents. Another advantage is that the well-characterized biology of the mouse and availability of commercially produced murine reagents will likely support future developments and expanded applications of this assay. The IWG noted that there may be regulatory situations that require chemicals to be tested in a clinically- or commercially relevant solvent or product formulations. It was noted that some testing situations will require continued use of the traditional GP test methods. These include the evaluation of metals or metal compounds, strong irritants, and laboratory situations where use of the LLNA is not feasible due to its current requirement for the use of radioisotopes. The IWG acknowledged that, with future development efforts, the LLNA holds great potential for expanded applications and use. ICCVAM endorsed the IWG and Panel recommendations and forwarded them to the Director, NIEHS, who in turn sent them to the participating ICCVAM agencies for regulatory acceptance consideration.

11. Acceptance of the LLNA by Federal Agencies

The U.S. Environmental Protection Agency (EPA), Food and Drug Administration (FDA), and Occupational Safety and Health Administration (OSHA) announced their acceptance of the LLNA as a stand-alone test method at the October 1999 meeting of the NTP Advisory Committee on Alternative Toxicological Methods (ACATM). They agreed that the LLNA can be used to assess the ACD potential of chemicals and products when used in accordance with the recommendations in the LLNA Panel Report. The acceptance process for each agency varied often due to the specific nature of the products and agents regulated by each agency.

U. S. EPA Consideration of the LLNA

The U.S. EPA accepted the ICCVAM position on the validity of the assay and supports its use as the first option for the identification of chemicals with contact sensitizing potential. U. S. EPA guidelines for the evaluation of chemicals for contact hypersensitivity potential are currently being revised to incorporate the LLNA as an acceptable method that can be used to meet U. S. EPA requirements. The U.S. EPA plans to identify situations in which the LLNA should be the recommended primary test, and situations in which the assay should not be recommended because of limitations relating to particular agents and regulatory situations.

FDA Consideration of the LLNA

The FDA reviewed the LLNA on a Center-by-Center basis. The Center for Drug Evaluation and Research (CDER) concluded that the LLNA is an acceptable stand-alone alternative to the GPMT for hazard

identification of dermally applied contact sensitizing agents within the limitations acknowledged by the Panel. The Center for Biologics Evaluation and Research (CBER) concluded that the LLNA is an acceptable alternative to the GPMT. CBER proposes to recommend the LLNA when skin sensitization data are necessary and appropriate. The Center for Veterinary Medicine (CVM) concluded that the LLNA is an acceptable alternative, but has no need to include the method in its testing guidelines. On the rare occasions when a veterinary drug is tested for hypersensitivity potential, preference is given to the use of the most appropriate test species. The Center for Devices and Radiological Health (CDRH) concluded that the LLNA is acceptable with qualifications similar to those noted by the Panel. Additionally, CDRH indicated that proper positive and negative controls for testing extracts of materials/devices should be included in the testing paradigm. The Center for Food Safety and Applied Nutrition (CFSAN) concluded that the LLNA is a valid, stand-alone alternative to the GPMT for detecting ACD mediated by food and color additives and cosmetics, within the limitations expressed by the IWG.

OECD Adoption of the LLNA

The LLNA was adopted as Test Guideline 429 in April 2002 by the Test Guidelines Programme of the Organisation for Economic Co-operation and Development (OECD). The draft guideline was distributed for review and commentary by the 30 member countries of the OECD in November 2000. The ICCVAM Panel Report on the LLNA was distributed with the proposed guideline to substantiate its scientific validity.

12. Implementation

ICCVAM, in conjunction with the International Life Sciences Institute (ILSI) Health and Environmental Sciences Institute's (HESI) Alternatives to Animal Testing Technical Committee, organized a one-and-a-half day training workshop on the LLNA in January 2001. The purpose of the workshop was to familiarize the regulated community with the LLNA, and the manner in which agencies expect the method to be conducted and interpreted. The workshop was designed for scientists from industry, government, and academia who had an interest in learning more about the LLNA. The primary objective of the workshop was to assist these participants in gaining a practical understanding of the theory and application of the LLNA.

The workshop provided information on the scientific basis of the assay and its development and validation. Practical conduct of the assay was emphasized including instruction on standard protocols and data interpretation. A case study exercise provided workshop participants an opportunity to evaluate data and to apply knowledge gained from the lecture. This exercise allowed for discussion of specific scientific and regulatory issues and identification of assay advantages and limitations. Constructive dialogue between the industry and regulatory agency staff at this workshop should expedite the implementation and global acceptance of the LLNA.

The rigorous and comprehensive review provided by the ICCVAM process afforded government regulatory agencies reasonable assurance of the usefulness and limitations of this new method.

13. REFERENCES

Basketter, D.A., and Scholes, E.W. (1992). Comparison of the local lymph node assay with the guinea pig maximization test for the detection of a range of contact allergens. Food Chem. Toxicol. 30, 65-69.

Basketter, D.A., Scholes, E.W., and Kimber, I. (1994). The performance of the local lymph node assay with chemicals identified as contact allergens in the human maximization test. Food Chem. Toxicol. 32, 543-547.

Basketter, D.A., Gerberick, G.F., Kimber, I., and Loveless, S.E. (1996). The local lymph node assay—A viable alternative to currently accepted skin sensitisation tests. Food Chem. Toxicology 34, 985-997.

Basketter, D.A., Scholes, E.W., Kimber, I., Botham, P.A., Hilton, J., Miller, K., Robbins, M.C., Harrison, P.T.C., and Waite, S.J. (1991). Inter-laboratory evaluation of the local lymph node assay with 25 chemicals and comparison with guinea pig test data. Toxicol. Methods 1, 30-43.

Chamberlain, M., and Basketter, D.A. (1996). The local lymph node assay: status of validation. Food Chem. Toxicol. 34, 999-1002.

Gerberick, G.F., House, R.V., Fletcher, E.R., and Ryan, C.A. (1992). Examination of the local lymph node assay for use in contact sensitization risk assessment. Fundam. Appl. Toxicol. 19, 438-445.

ICCVAM (Interagency Coordinating Committee on the Validation of Alternative Methods). (1999). The Murine Local Lymph Node Assay: A test method for assessing the allergic contact dermatitis potential of chemicals/compounds. NIH Publication No. 99-4494, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina.

ICCVAM. (Interagency Coordinating Committee on the Validation of Alternative Methods). (1997). Validation and Regulatory Acceptance of Toxicological Test Methods: A Report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods. NIH Publication 97-3981. National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina. Available on the Internet at http://ntp-server.niehs.nih.gov/htdocs/ICCVAM/iccvam.html.

Kimber, I., and Weisenberger, C. (1989). A murine local lymph node assay for the identification of contact allergens. Assay development and results of an initial validation. Arch. Toxicol. 63, 274-282.

Kimber I., Hilton, J., and Botham, P.A. (1990). Identification of contact allergens using the murine local lymph node assay: Comparisons with the Buehler occluded patch test in guinea pigs. J. Appl. Toxicol. 10, 173-80.

Kimber, I., Hilton, J., Botham, P.A., Basketter, D.A., Scholes, E.W., Miller, K., Robbins, M.C., Harrison, P.T.C., Gray, T.J.B., and Waite, S.J. (1991). The murine local lymph node assay: Results of an inter-laboratory trial. Toxicol. Lett. 55, 203-213.

Kimber, I., Hilton, J., Dearman, R.J., Gerberick, G.F., Ryan, C.A., Basketter, D.A., Lea, L., House, R.V., Ladics, G.S., Loveless, S.E., and Hastings, K.L. (1998). Assessment of the skin sensitization potential of topical medicaments using the local lymph node assay: An inter-laboratory exercise. J. Toxicol. Environ. Health 53, 563-579.

Kimber, I., Hilton, J., Dearman, R.J., Gerberick, G.F., Ryan, C.A., Basketter, D.A., Scholes, E.W., Ladics, G.S., Loveless, S.E., House, R.V., and Guy, A. (1995). An international evaluation of the murine local lymph node assay and comparison of modified procedures. Toxicology 103, 63-73.

Kimber, I., Hilton, J., and Weisenberger, C. (1989). The murine local lymph node assay for identification of contact allergens: A preliminary evaluation of in situ measurement of lymphocyte proliferation. Contact Derm. 21, 215-220.

Landsteiner, K., and Jacobs, J. (1935). Studies on sensitization of animals with simple chemical compounds. J. Exp. Med. 61, 643-656.

Loveless, S.E., Ladics, G.S., Gerberick, G.F., Ryan, C.A., Basketter, D.A., Scholes, E.W., House, R.V., Hilton, J., Dearman, R.J., and Kimber, I. (1996). Further evaluation of the local lymph node assay in the final phase of an international collaborative trial. Toxicology 108, 141-152.

Marzulli, F.N., and Maibach, H.I. (Eds.). (1996). Dermatotoxicology, 5th Ed. Taylor & Francis, Washington, D.C.

Sailstad, D., Hattan, D., Hill, R., and Stokes, W. (2000). The first method evaluation by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM): The murine local lymph node assay (LLNA). In Progress in the Reduction, Refinement and Replacement of Animal Experimentation (M. Balls, A.-M. van Zeller, and M.E. Halder, Eds.), pp. 425-433. Elsevier Science, Amsterdam.

Scholes, E.W., Basketter, D.A., Sarll, A.E., Kimber, I., Evans, C.D., Miller, K., Robbins, M.C., Harrison, P.T.C., and Waite, S.J. (1992). The local lymph node assay: Results of a final inter-laboratory validation under field conditions. J. Toxicol. Appl. Pharmacol. 12, 217-222.

Selgrade, M.J.K., Germolec, D.R., Luebke, R.W., Smialowicz, R.J., Ward, M.D., and Sailstad, D.M. (2001). Immunotoxicity. In Introduction to Biochemical Toxicology, 3rd Edition (E. Hodgson and R.C. Smart, Eds.), pp. 597-561. J. Wiley and Sons, NY, NY.

Stokes, W., and Hill, R. (2000). The role of the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) in the evaluation of new toxicological testing methods. In Progress in the Reduction, Refinement and Replacement of Animal Experimentation (M. Balls, A.-M. van Zeller, and M.E. Halder, Eds.), pp. 385-394. Elsevier Science, Amsterdam.

Sailstad, D., Hattan, D., Hill, R., and Stokes, W.S., Evaluation of the Murine Local Lymph Node Assay (LLNA) I: The ICCVAM Review Process. Regul. Toxicol. Pharmacol. 34(3):249-257, 2001.

Dean, J., Twerdok, L., Tice, R., Sailstad, D., Hattan, D. and Stokes, W.S., Evaluation of the Murine Local Lymph Node Assay (LLNA) II: Conclusions and Recommendations of an Independent Scientific Peer Review Panel. Regul. Toxicol. Pharmacol. 34(3):258-273, 2001.

Haneke, K., Tice, R., Carson, B., Margolin, B., and Stokes, W.S., Evaluation of the Murine Local Lymph Node Assay (LLNA): III. Data Analyses Completed by the National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM). Regul. Toxicol. Pharmacol. 34(3):274-286, 2001.

USC (United States Code) NIH/National Institutes of Health Revitalization Act. Public Law 103-43. 42 USC Section 205. Washington, DC: US Government Printing Office, 1993.

ICCVAM Authorization Act of 2000, Public Law 106-545, 106th Cong., 2nd sess. (December 19, 2000).

Example 5

The Up-and-Down Procedure for Acute Oral Toxicity (TG 425)

1. Introduction

A US team of toxicologists and statisticians developed an improved version of the up-and-down procedure as one of the replacements for the traditional acute oral toxicity test formerly used by OECD Member countries to characterize industrial chemicals, pesticides and their mixtures. This method, OECD Test Guideline 425 (2001), for applications that use the LD50, achieves significant reductions in animal use. It employs sequential dosing, together with use of computer-assisted computational methods during the execution and calculation phases of the test. Staircase design, a form of sequential test design, can be applied to acute toxicity testing with its binary experimental endpoints (yes/no outcomes). The staircase design used in the Test Guideline builds in efficiencies by concentrating most doses near the region of the LD50, and by using flexible or adaptive stopping rules to accommodate a variety of chemicals. The up-and-down procedure provides a point estimate of the LD50 and approximate confidence intervals, in addition to observed toxic signs, for the substance tested. However, when assumptions about standard deviation of the test population diverge significantly from actual values, point estimates of the LD50 may not be possible. In these cases, an LD50range estimate is provided. The test method does not provide information about the dose-response curve.

2. History

In light of the heavy use of animals in traditional test methods, R.D. Bruce (1985) applied the up-and-down procedure, a sequential test design, to acute toxicity testing. The test design assumed a value for sigma, the standard deviation of the curve, which is suitable for certain industrial chemicals.

The Up and Down technique was evaluated in laboratory studies in 1987. Subsequently, two other studies were done comparing the results of the up-and-down procedure to those using traditional LD50 test methods, for 35 substances (Lipnick et al.,1995)

The goals of the United States design team were to fulfil the OECD mandate to develop a robust test that would be applicable to a wide variety of pesticides, industrial chemicals, and their products. The test should have the ability to be performed, and for results to be calculated, without the need to assume a value for the standard deviation (SD). Computer simulations were used to evaluate the performance of the previous version of the OECD Test Guideline (OECD TG 425, 1998) and to determine appropriate changes to optimize the method's performance without actually testing animals in the laboratory.

3. Rationale for Experimental Design Used in Test Guideline 425

Sequential test designs, in contrast to test designs with fixed sample size involving replicate sampling, can achieve efficiencies in the number of test samples needed by sampling one or a few test subjects at a time until just enough measurements are made to evaluate the experimental endpoint of concern with the desired precision. In general, sequential designs can be applied to either qualitative (yes/no) or quantitative outcomes, but only the former are considered in this paper. Staircase designs permit trials to rapidly converge on the region of interest, such as the ED50 or LD50 in a toxicity test.

For suitable applications in toxicology, the use of sequential design and non-traditional calculation methods can lead to reduction of animal usage, while maintaining the ability of the test to measure desired experimental results. Acute toxicity testing, which measures the adverse effects that occur within a short

time of administration of single dose of a chemical, is one such candidate for the application of sequential design.

4. Test Description

The up-and-down procedure for acute oral toxicity involves testing young adult animals one at a time in a staircase fashion. Unless the conditions for terminating the test have been reached, the next dose level depends on the results in the previous animal. If that animal survived, then the next dose is higher, while if the last animal died, then the next dose is lower. This permits the experiments to converge on the region of the LD50, which is the inflection point corresponding to the median response of the log-normal dose-response curve.

The OECD TG 425 (2001) uses sequential dosing and calls for testing to be performed in a single sex to reduce variability in the test population. Animals are tested individually in a staircase fashion using a dose progression and starting dose based on characteristics of the chemical being evaluated. A flexible or adaptive stopping rule limits the number of animals in the main test, while allowing the method to be applied to chemicals with a wide range of slopes of the dose-response curve. A computer program is available to determine if conditions for terminating the test have been reached. Guidance is provided for use of all available information to determine initial dosing and dose progression spacing. Initial doses should be set at sub-lethal levels and dose progression based on approximate slope. However, the TG includes recommendations for default values for starting dose and dose progression for use in the absence of initial information. As in the traditional acute test, this replacement up-and-down procedure provides for clinical observations over 14 days. While the statistical procedures are more complicated than those for a non-sequential test, much of the complexity can be handled by computer software.

In addition, a sequential limit test using up to five animals has been substituted for the ten-animal batch limit test. Sequential observations permit a separate decision about the contribution of each animal to the determination of whether the LD50 lies above or below the limit dose.

5. Use of Statistical Simulation for Development and Validation of the Up-and-Down Procedure

By varying assumed values of the LD50 and the SD for a hypothetical set of animals exposed to a toxic chemical, it is possible to have a computer generate responses from a randomly chosen sample of the hypothetical population. A computer can use these hypothetical responses to simulate the results from thousands of small samples of the underlying population. Since the underlying mean and SD of the test population is known, these simulations then can be used to determine if changes in the test design would improve the ability of the up-and-down procedure test to estimate the mean and SD. By varying the SD assigned to the hypothetical set of animals, it is possible to simulate the degree of variation in the experimental sample's response that would occur because of animal-to-animal, inter-, and intra-laboratory species, strain, sex, age, and housing variability. Such simulations have shown that the new sampling technique used in the up-and-down procedure has a much better chance of placing the estimated LD50 close to the value that was used in defining the underlying hypothetical population even when the starting dose is inappropriate. This type of comparison would not be practical using actual animal tests, since it would be impossible to determine whether each small sample tested is providing correct or incorrect estimates of the underlying population.

Actual animal testing was not necessary for determination of the validity of the new statistical design. It is not appropriate or possible to compare changes in sampling technique or to assess the ability of a new statistical design to accurately estimate the mean and SD of the population based on the results of a few runs of a test. This is because there is no way to determine the goodness of fit of the statistical procedure from a few samples. However, computer simulations can be used to compare the results of thousands of

individual hypothetical tests. By using a large series of such simulations varying starting dose, dose progression and assumptions about mean number and SD of the underlying population, it is possible to test how often a new statistical sampling technique will accurately estimate the LD50 and SD of the population.

Simulations showed that the number of test subjects needed to provide an acceptable degree of accuracy depends, in part, on the slope of the dose-response curve of the test population. However, in most cases, the slope is not known in advance of testing. Therefore, in order to allow the up-and-down method to be applied to a wide variety of chemicals with reasonable reliability, a flexible stopping rule, using criteria based on an index related to the statistical error, was developed and incorporated into the test. For chemicals with higher slopes, the stopping rules will be satisfied with four animals after the first reversal. Additional animals may be needed for chemicals with dose-response slopes below 4. In the interest of animal welfare, testing in any case is terminated at 15 animals. Simulations have shown that the average animal use is expected to be seven to nine animals per test.

6. Performance of the Revised TG 425

The revised OECD Test Guideline 425 (OECD 2001) has improved performance for prediction of the point estimate of lethality (LD50) and confidence intervals for chemicals with wide variability of response characteristics, even when the approximate LD50 and dose-response slope are not known. The efficiencies of the new up- and-down dosing routine allow the LD50 to be estimated using relatively few animals. However, the Test Guideline does not provide for determination of the slope of the dose-response curve.

Although flexible stopping rules allow the up-and-down procedure to be applied to test materials with a wide range of slopes, for optimum performance of the up-and-down procedure, the dose progression used should be based on an accurate estimate of the SD. In addition, to account conservatively for any bias in the LD50 estimate, it is essential that dosing be initiated below the actual LD50. Setting initial doses at sublethal levels also ensures that LD50 values are not underestimated while reducing distress in the animals.

In traditional LD50 tests, the stated confidence interval, for example 95%, is exactly what is calculated. However, for the up-and-down procedure, the algorithm used to compute 90%, 95%, or 99% profile likelihood confidence intervals is not exact but approximate, so that in some situations, the stated confidence interval will not provide the desired coverage or may provide more than the desired coverage. Simulations indicate that actual coverage for the nominal 95% confidence interval falls below 95% when the slope is shallow and above 95% when slopes are very steep. The nominal 95% confidence interval will have coverage of at least 90% if the slope is 2-4 or greater (SD 0.25-0.5 or smaller). For most situations, the coverage will be better than 90% if the slope is 2 or greater. Coverage will be 80% or better if the slope is at least 1. For slopes as low as 0.5, the coverage may be as low as 70%.

The up-and-down procedure does not allow for characterization of the slope since efficient management of doses toward the region of the LD50 does not provide enough doses in the wings of the dose-response curve. If slope is needed, another method would be used. The revised up-and-down procedure Test Guideline can be used for chemicals with a wide variety of actual dose-response slopes and can be used for hazard classification and certain other hazard and risk assessment purposes. The five-animal sequential plan used in the limit test produces results that are almost as good as those for the present ten-animal fixed sample plan, while averaging three to five animals per test. That represents a substantial reduction of animal subjects over the ten-animal fixed sample plan.

7. Peer Review

The revised Test Guideline was peer reviewed by independent panels of experts convened by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM 2001) and by the EPA Scientific Advisory Panel (EPA 2001).