Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development 24-Jul-2008

English - Or. English

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

Cancels & replaces the same document of 21 July 2008

SERIES ON TESTING AND ASSESSMENT
Number 43

GUIDANCE DOCUMENT ON MAMMALIAN REPRODUCTIVE TOXICITY TESTING AND ASSESSMENT

JT03249193

ENV/JM/MONO(2008)16
Unclassified

English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 43

GUIDANCE DOCUMENT ON MAMMALIAN REPRODUCTIVE TOXICITY TESTING AND ASSESSMENT

IOMC

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

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

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris 2008

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, revised 2006)*

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

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

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

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

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

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

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

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

No. 10, *Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data (1998)*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

No. 35, *Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies (2002)*

No. 36, *Report of the OECD/UNEP Workshop on the use of Multimedia Models for estimating overall Environmental Persistence and long range Transport in the context of PBTS/POPS Assessment (2002)*

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

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

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (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 Mammalian Reproductive Toxicity Testing and Assessment (2008)*

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

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

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

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

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

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

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

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

No. 52, *Comparison of emission estimation methods used in Pollutant Release and Transfer Registers (PRTRs) and Emission Scenario Documents (ESDs): Case study of pulp and paper and textile sectors (2006)*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

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

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

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) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This *Guidance Document (GD) on Mammalian Reproductive Toxicity Testing and Assessment* is intended to provide guidance on methodological aspects and interpretation of data in the testing of chemicals for potential human and other mammalian reproductive toxicity. It is also covering the relationship with neurotoxicity testing. The document constitutes an essential supplement to existing OECD Test Guidelines which can be used to obtain information on the potential reproductive toxicity of chemicals.

The project for a guidance document on reproductive toxicity testing and assessment was included in the work plan in 1996. After several expert meetings, the first version of the draft guidance document on reproductive toxicity testing and assessment was circulated to the Working Group of the National Coordinators of the Test Guidelines Programme (WNT) for comments in November 2004.

To address the comments received and revise the draft GD, a drafting group in charge of the GD had a last meeting in October 2005, in Paris. The GD was revised and a new draft was circulated in December 2007 to the WNT for comments. A final draft GD, prepared by the Secretariat in February 2008 on the basis of the comments received, was agreed by the WNT at its 20th meeting in April 2008. The WNT also agreed that this GD would be revised when new Test Guidelines for the testing of potential reproductive toxicity of chemicals would be available.

This document is published on the responsibility of the Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology.

Contact for further details:
Environment, Health and Safety Division
Environment Directorate
Organisation for Economic Co-operation and Development
2, rue André-Pascal
75775 Paris Cedex 16, France
E-mail : env.edcontact@oecd.org

TABLE OF CONTENTS

I. INTRODUCTION

History of the document

1. The OECD Working Group on Reproduction and Developmental Toxicity met in Copenhagen, Denmark in June 1995. The meeting recognised that there was a need for a Guidance Document on developmental and reproduction toxicity testing, covering testing strategies, approaches for tier testing, relationship with neurotoxicity testing and data interpretation procedures. The 7th WNT meeting approved to develop the Guidance Document in October 1996 and the US EPA volunteered to take the lead in this activity.

2. An expert group met several times to co-ordinate the drafting of the document. In June 2003, the US EPA hosted a Meeting for the expert group in Washington DC to resolve the last problems and assign writing tasks. The first version of Guidance Document No.43 was circulated to Member countries for comments in November 2004. The document was generally well received, however, a considerable number of comments were provided. Many comments were of purely editorial nature but a number of more complex issues were also raised, therefore the Secretariat invited the expert group to Paris in October 2005 to address these issues and revise the draft Guidance Document accordingly.

Purpose of the document

3. This document is intended to provide guidance on methodological aspects, interpretation of data and an overall approach for testing of chemicals for potential human and other mammalian reproductive toxicity.

4. The document constitutes an essential supplement to existing OECD Test Guidelines that can be used to obtain information on the potential reproductive toxicity of chemicals. Specific OECD Test Guidelines include the one- and two-generation toxicity study (TG 415 and 416), prenatal developmental toxicity study (TG 414), developmental neurotoxicity study (TG 426) and the reproduction/developmental toxicity screening tests (TG 421 and 422). However, data from other toxicity studies *e.g.,* repeated dose toxicity studies for systemic toxicity (TG 407, 408 and 409) may indicate potential reproductive toxicity and should be considered in the assessment as well as existing human data.

5. The term Endocrine Disruption (ED) denotes the ability of exogenous chemicals to alter function(s) of the endocrine system and consequently cause adverse health effects in an intact organism, or its progeny, or (sub) populations. Extensive information has become available on ED in the recent years (CSTEE 1999). OECD has initiated a specific programme, the Task Force on Endocrine Disrupters Testing and Assessment (EDTA) to develop new Test Guidelines, and to revise existing ones for detection of ED, presently focusing on the reproductive system. The current Test Guidelines may enable the identification of ED-related reproductive effects if conducted with specific attention to this issue. However, current Test Guidelines for toxicological testing could be improved with the aim of increasing sensitivity and specificity of detection of ED-related effects. This Guidance Document will include considerations on ED-related effects but ED is not the main focus of the document.

6. A number of organisations in several countries have written documents on the evaluation of reproductive toxicity data, *e.g.,* ECETOC 1983, 1992, 2002; EHC 1984; LST 1989; NIOH (AMI) 1994;

DEPA 1995; IPCS-WHO 2001 and OECD 2002a. Annex 3 gives guidance on an approach to hazard characterisation with regard to reproductive toxicity.

Definition of developmental/reproductive toxicity and principles of reproductive toxicity hazard assessment

7. Reproduction is the biological process that ensures the continuation of the species. Existing genetic material is passed to the next generation.

8. Reproductive toxicity includes adverse effects on sexual function and fertility in adult males and females as well as developmental toxicity in the offspring. The distinction between developmental and reproductive toxicity is somewhat arbitrary in that developmental exposures can result in effects on reproduction, and vice versa.

9. Adverse effects on reproductive ability or capacity include alterations in the female and male reproductive organs or related endocrine systems *i.e.,* effects on:

- onset of puberty;
- gamete production and transport;
- reproductive cycle normality;
- sexual behaviour;
- fertility;
- parturition;
- premature reproductive senescence, or;
- modifications in other functions that are dependent on the integrity of the reproductive systems, *e.g.,* lactation.

10. Developmental toxicity taken in its widest sense includes any effect which interferes with normal development of the conceptus, either before or after birth, and resulting from exposure of either parents prior to conception, or exposure of the developing offspring during prenatal development, or post-natally, to the time of sexual maturation. These effects can be manifested at any point in the life-span of the organism. The major manifestations of developmental toxicity include:

- death of the developing organism;
- structural abnormality;
- altered growth (including maturation), and;
- functional deficiency.

(See the following references: Clark (1999), Cooper and Goldman (1999), ECETOC (1983), EHC (1984), Heindel (1999)).

11. Reproductive toxicity may result from single or repeated doses of a chemical. Since both single and repeated exposures are possible scenarios for human exposure, the reproductive toxicity assessment should ideally consider both situations

12. The assessment of reproductive toxicity will be discussed in the sections for specific endpoints. The available methodologies, normal range of variation for test species, interrelationship among endpoints, relationship of endpoint to functional outcome, differences among test species, relationship with maternal toxicity, reversibility of outcome, and relevance to humans will be discussed in the sections on prenatal endpoints, postnatal endpoints, and adult reproductive toxicity.

Identification and assessment of the effects of chemicals on reproduction and development

Reproductive Toxicity Tests

13. The reproductive toxicity tests constitute an important part of the toxicological testing program in the health assessment of chemicals. The data from the toxicological testing shall provide the necessary information with respect to:

- Hazard identification
- Dose-effect assessment (if possible estimation of a NOAEL or benchmark dose*)
- Extrapolation (prediction of adverse effects in other species, particularly in humans)
- Prediction of safe levels of exposure in humans

* The present OECD Test Guidelines are not optimally designed for a benchmark dose approach.

14. Tests for reproductive toxicity should ideally be able to identify one or more of the following effects on reproduction:

- Impairment of male or female reproductive functions or capacity, *i.e.,* adverse effects on libido, oestrus cycle, sexual behaviour, any aspect of spermatogenesis or oogenesis, or hormonal activity or physiological response which would interfere with the capacity to fertilise, fertilisation itself or development of the fertilised ovum up to and including implantation.

- Induction of non-inheritable harmful effects on the progeny, *i.e.,* in the widest sense, any effect interfering with normal development, both before and after birth should be included. Morphological malformation(s) and functional disturbances (*e.g.,* hormonal reproductive effects, neurological) should be evaluated.

15. Many different experimental methods for investigating toxic effects of chemicals on reproduction and development are in use (OECD 2002b; Toppari *et al.,* 1996; ECETOC 2002; Meyer and Svendsen 2003). Several tests are standardised and Test Guidelines have been issued by various governmental agencies and international organisations. The following sections will focus mainly on standardised and accepted regulatory OECD Test Guidelines, see table 1. Tests other than those included in table 1 can reveal effects which indicate a potential of a chemical to interfere with normal reproduction, *e.g.,* the dominant lethal test, fertility assessment by continuous breeding, and repeated dose toxicity testing where the gonads are subjected to pathological examination.

16. A new OECD Test Guideline on an extended one generation reproductive toxicity study is currently under development. This Test Guideline will be largely based on the International Life Science Institute (ILSI)-Health and Environmental Science Institute (HESI), Agricultural Chemical Safety Assessment (ACSA) Technical Committee proposed life stage F1 extended study, published in Cooper *et al*. (2006). Several improvements and clarifications will be made to the study design. To meet the different regulatory needs and to address the common goal of reduced/refined animal testing, this TG will provide flexibility and a more efficient testing approach by stressing the importance of starting with existing knowledge and using in-life observations to guide and tailor the testing. The TG will consist of three cohorts. In addition to the main reproductive/developmental endpoints (cohort 3), the TG will offer the possibility of assessing developmental, neurotoxicity and immunotoxicity (cohorts 1&2 respectively). The conduct of cohorts 1&2 will depend upon the nature of the existing chemical database and needs of different regulatory authorities. The production of a second generation will be contingent on existing data and observations during the in-life portions of the study. In this context Janer *et al*. (2007) made a retrospective analysis evaluating 176 multigeneration studies to assess potential differences between the first and the second

generation. The results of this analysis support the proposal of replacing the current two-generation study, by an extended one generation study. Thus this new TG could be an alternative to TG 415 and 416. As this TG is currently under development, it has not been included in this Guidance Document, but will be in the next edition of the Guidance.

Table 1. Overview of *in vivo* tests for reproductive toxicity testing

Test	Exposure period	Endpoints in parental and/or offspring	Guideline(s)
Generation studies	Continuously over one, two or several generations	Growth, development and viability. Pregnancy length and birth outcome Histopathology of sex organs and target organs Fertility Oestrus cyclicity and sperm quality in TG 416	TG 415: One-generation Study TG 416: Two-generation Study
Prenatal Developmental Toxicity Study (Teratology study)	From implantation to the day before birth	Litter composition (*e.g.,* resorptions, live, dead foetuses) Embryonic development Foetal growth Morphological variations and malformations.	TG 414: Prenatal Developmental Study
Developmental Neurotoxicity Study	During pregnancy and lactation	Pregnancy length and birth outcome Physical and functional maturation Behavioural changes due to CNS and PNS effects Brain weights and neuropathology	TG 426: Developmental Neurotoxicity Study
Reproduction/ Developmental toxicity screening test	From 2 weeks prior to mating until day 4 post-natally	Fertility Pregnancy length and birth outcome Histopathology of sex organs and target organs (and brain in TG 422) Foetal and pup growth and survival until day 3	TG 421 and 422

17. During recent years many *in vitro* test systems have been proposed as alternatives to animal testing for developmental toxicity (ATLA 2002). These tests usually address single events of the reproductive cycle and are therefore insufficient for the assessment of adverse *in vivo* effects and they do not replace animal testing in the risk assessment of chemicals. However, these tests may be useful for screening of closely related chemicals and for elucidating the mechanisms underlying the effects. They may also be essential elements of stepwise testing and assessment strategies (see Appendix 3). Work is also conducted on non mammalian species, as an example, the US NTP is working on a soil nematode (*Caenorhabditis elegans)* model, as a medium throughput, non-mammalian test for developmental and neurological toxicity (US NTP, 2006). With respect to other in vivo tests, see paragraphs 44-45.

Assessment of Systemic Effects in Adult Animals

18. Repeated dose toxicity testing in adult animals provides information on the potential for systemic toxicity by investigations of growth, clinical symptoms, behaviour, haematology, biochemistry, organ weights, as well as pathology and histopathology of organs, including sex organs. Changes in the histopathology of certain reproductive organs (e.g., testes, ovaries, etc.) may be sensitive indicators of a potential reproductive hazard, and positive effects on these parameters observed in general toxicity studies may be valuable for the assessment of reproductive toxicity (Dent 2007; Chapin *et al*., 1998).

19. Reproductive toxicity testing can provide information on a number of developmental effects, such as malformations, growth retardation, embryo/foetal and postnatal death, fertility and functional effects on the CNS. Some of the Test Guidelines provide unique information, for example information on developmental effects on fertility and sex organs is only provided by the two-generation study, and effects on brain development and functions are investigated in the developmental neurotoxicity study.

20. The investigations of systemic effects in reproductive toxicity tests differ from those of repeated dose toxicity studies in adults, since *e.g.,* haematology and biochemistry are not usually investigated. In addition, the investigations of organ weight, pathology and histopathology are limited to the sex organs and identified target organs. Consequently, systemic effects induced during pre- or postnatal development (*e.g.,* on liver and kidneys) may not be identified in a standard reproductive toxicity study, unless a target organ was identified.

Significance of Experimental Data and their Relevance to Humans (Extrapolation)

21. The ultimate proof that a substance is a human reproductive toxicant can come only from information on the consequences of human exposure. This statement is valid regarding toxicological effects in general, as stated by Calabrese (1983):

> *"Cellular structure and biochemistry are remarkably alike across the entire animal kingdom, starting with the lipoprotein cell membrane, which affects the absorption of xenobiotics into the cell to metabolic processes like glycolysis, the Krebs' cycle, and numerous other aspects of intermediary metabolism. The similarity among animals on the cellular level is so apparent that it serves as the basis upon which scientists have extrapolated or inferred functions from one species to another"*

It is known that fundamental developmental processes are highly conserved within mammals. Therefore, the findings in experimental studies in laboratory animals, demonstrating developmental toxicity are indicative of a potential human response (Jakobsen and Meyer 1989; Schardein *et al.,* 1985; IPCS 2001). However, interspecies variations exist and the absence of developmental toxicity in one mammalian species may not always be indicative of an absence of toxicity in humans. These species and strain differences include variations in the absorption, distribution, metabolism and excretion of chemicals; in placental structure, permeability and blood flows (Schröder 1995); and in the genetic backgrounds of different species (Kawakami *et al*., 2006).

22. No single laboratory test species can be used to routinely predict the true human response to a given chemical. Tests in multiple species may increase the predictive reliability of laboratory animal test data; specific differences between the test species and humans should be considered in evaluating the relevance of particular effects to humans (Jakobsen and Meyer 1989).

23. Effects on fertility in rodents seem to be a good indicator for effects in humans, and most work on contraceptive agents in humans stems from original studies in rodents (Barlow and Sullivan 1982). However, it should be emphasized that sperm count in rodents must be drastically reduced before an effect on fertility is seen (see chapters 181-183). Since most of the reproductive studies are undertaken in litter-bearing animals, significant, but modest reductions in litter size may also be indicative of a functional deficit that can be related to sperm parameters (see chapter 194).

Single and Multi-Generation Studies

24. Test Guidelines for carrying out single and multi-generation studies have been published by the OECD and other organisations.

25. The purpose of these studies is to examine successive generations to identify possible effects of a substance on fertility of male and female animals; pre-, peri-, and post-natal effects on the ovum, embryo, foetus, and progeny, including teratogenic effects and peri- and post-natal effects on the mother.

26. The various Test Guidelines have a number of common requirements. The preferred species are the rat and the mouse. Other species may be used if relevant (*e.g.,* if differences in toxicokinetics between the preferred species and man are expected, to clarify ambiguous results or to further study observed effects). The test substance is administered to groups of animals (number of animals per group sufficient to yield about 20 pregnant animals at or near term). In general, the chemical is administered over at least one spermatogenic cycle and the last stages of oocyte maturation before the parent generation animals are mated. The exposure of the females is continued throughout the mating period and gestation up to termination after weaning of their litters. At least three treatment groups and a control group (untreated or vehicle-dosed) are to be used, unless a limit test is applied. Standardisation of litter sizes is optional (see chapter III). Ideally, unless limited by the physico-chemical nature or biological effects of the test substance, the highest dose level should induce toxicity but not mortality in the parental animals. The low dose should ideally not induce any observable adverse effects on the parents or offspring.

27. Birth weight, postnatal growth and survival are recorded, but it is important in the evaluation of the data to consider variations due to different litter sizes or different sex distribution in the litters as well as different postcoital age at birth (*i.e.*, duration of pregnancy length). A change in offspring body weight or weight gain is a sensitive indicator of developmental toxicity, in part because it is a continuous variable. In some cases, weight differences in offspring may be the only indicator of developmental toxicity in a generation study. While there is always a question remaining as to whether weight differences are a permanent or transitory effect, little is known about the long-term consequences of short-term foetal or neonatal weight changes. Therefore, weight differences should be considered as a relevant effect in establishing the NOAEL.

28. In the two-generation reproductive toxicity study (TG 416), assessment of effects on sperm quality and oestrus cyclicity in offspring has been included. In addition, the assessment of the offspring has been expanded with the recording of developmental milestones including some optional behavioural parameters and histopathology of sex organs, brain and identified target organs.

29. The multi-generation studies include exposure of both sexes during all stages of reproduction. It should provide a complete evaluation of the F1 animals, including their fertility and reproductive performance, which is not found in other studies. The F1 animals are unique considering that they can be regarded as an equivalent of an *in utero* derived repeated dose toxicity study.

30. The multi-generation studies are together with the developmental neurotoxicty study, the only toxicological tests which for rodents cover the early period of the development of the individual from the time of weaning up to about an age of five to six weeks where the exposure in repeated toxicity tests normally starts. Thus these tests open for the assessment of all end-points, provided that the parameters used are sufficient sensitive. However, these tests are labor-intensive and many scientific comments claim that the Guidelines for these tests demand too many parameters which could interfere with the validity of the data. Other scientific comments claim for more parameters. All together Guidelines for these reproductive tests still have undergo some revision e.g. introduction of new parameters and refinements or replacement of some of those already in. Endpoints such as neonatal death and malformations constitute a specific problem, since the commonly used laboratory animals may eat dead or seriously malformed pups immediately after birth. An effect may, therefore, only be indicated indirectly by a smaller litter size. If only a few pups were malformed or dead, the reduction in litter size will be small compared to the normal variation in litter size and may therefore go undetected or not reach statistical significance. Pre-implantation losses and resorptions are indicated in an indirect way as a decreased litter size and the sensitivity for these effects may be rather low.

Prenatal Developmental Toxicity Study, Teratology Study

31. The prenatal developmental toxicity study (TG 414) is the method for examining embryo-foetal toxicity as a consequence of exposure during pregnancy. In the past, there was a tendency to consider only malformations and death as relevant endpoints in teratology studies. Today the test focuses on growth retardation, structural abnormalities and lethality. Consequently, the title of the TG 414 has been changed to "Prenatal Developmental Toxicity Study".

32. The preferred species include rodents (*e.g.,* the rat and mouse) and non-rodents (*e.g.,* the rabbit). Other species may be used if relevant (*e.g.,* if differences in toxicokinetics between the preferred species and humans are expected, to clarify ambiguous results or to further study observed effects).

33. Young mature virgin females are artificially inseminated or mated with males. The time of mating is established by observation of mating (*e.g.,* rabbits), identification of a seminal plug (mixture of sperm, secretion from coagulation gland and cells and mucus from vagina), vaginal smear (in rats) or by noting the time of insemination (*e.g.,* for pigs and rabbits). According to the current OECD TG 414, three dose levels and a control group (untreated or vehicle control; the group size is 20 pregnant animals, however 16 as a minimum in TG 414) are used in order to establish a dose-effect relationship. Under the superseded TG 414 (1981), the pregnant female rats were exposed at least during the period of organogenesis, *i.e.,* between day six when implantation occurs, and day 15 (the corresponding periods for mice and rabbits are days 6-15 and days 6-18, respectively). This period has been found to be the most sensitive to the induction of malformations (the corresponding sensitive period for humans is between days 18-60 of pregnancy). However, development of *e.g.,* sex organs and brain continues after day 15 and consequently malformations of such organs may not be discovered if exposure is stopped on day 15. In the current TG 414 (2001), the dosing period extends from implantation to scheduled caesarean section. If preliminary studies, when available, do not indicate a high potential for pre-implantation loss, treatment may be extended to include the entire period of gestation, from mating to the day prior to scheduled kill. The animals are observed daily for clinical changes. Body weight and food consumption are recorded throughout the gestation. The day before anticipated birth, the uterus is removed by caesarean section and the uterus and the foetuses are examined. The dam is examined macroscopically for any structural abnormalities or pathological changes. If dosing is initiated before or at the time of implantation, the pre-implantation loss, *i.e.,* the number of embryos lost prior to implantation is evaluated.

34. The total number of implantations, *i.e.,* living foetuses, dead foetuses and resorptions (embryos that die early and are assimilated) are noted. The degree of resorption (*i.e.,* the extent to which the embryo has

been resorbed i.*e.,* total, early, and late resorptions) is recorded in order to establish the time of death of the embryo during the pregnancy.

35. The foetuses are sexed, weighed and examined for gross malformations. Retarded growth and effects on visceral and skeletal development are evaluated. Live pups removed during this procedure should be humanly killed prior to visceral/skeletal evaluation.

36. The Prenatal Developmental Toxicity Study (TG 414) is very suitable for the demonstration of intra-uterine death after implantation (resorptions). In studies where dosing is started before implantation, pre-implantation loss may also be assessed.

37. The foetal weight is assessed, but it is important to consider the influence of different litter sizes or sex distribution in the analysis of the data.

Developmental Neurotoxicity Studies

38. An OECD Test Guideline for a "Developmental Neurotoxicity Study" (TG 426) has been recently adopted, based upon the US EPA guideline (US EPA 1998a). Developmental neurotoxicity studies are designed to obtain data on the potential functional and morphological hazards to the nervous system arising in the offspring from exposure of the mother during pregnancy and lactation. These studies can identify changes in behaviour due to effects on the central nervous system (CNS) and the peripheral nervous system (PNS). As behaviour is affected by the function of other organs such as liver, kidneys and the endocrine system, toxic effects on these organs in offspring may also be reflected in general changes in behaviour. No single test is able to reflect the entire complex and intricate function of behaviour. For testing of behaviour a range of parameters is used to identify changes in individual functions. Methodology employed in developmental neurotoxicity studies is described in several reviews (Adams 1986; Francis *et al.,* 1990; IPCS 2001). The test methods may generally be grouped into tests of physical development, simple reflexes, motor function, sensory development and functions, spontaneous activity, learning and memory, and neuropathology.

39. The preferred species is the rat. The OECD TG recommends groups of 20 litters per dose level with dosing during gestation and lactation. After birth, the number of progeny is recorded and the litters may be adjusted so that each contains the same number of pups. In order to determine whether the tested chemical affects the offspring directly via the mother's milk or indirectly, either via a change in milk production or as a result of a change in the behaviour of the exposed mothers, cross-fostering may be employed. Cross fostering is a method where litters from exposed mothers are reared by control mothers and vice versa.

40. The evaluation of the offspring consists of observations to detect gross neurological and behavioural abnormalities, assessment of physical development, reflex ontogeny, motor activity, motor and sensory function, learning and memory, and evaluation of brain weights and neuropathology (including morphometric examinations) during postnatal development and adulthood.

41. A developmental neurotoxicity study can be conducted as a separate study or incorporated into a reproductive study *e.g.,* a one or two-generation study. When developmental neurotoxicity study is incorporated within or attached to another study, it is imperative to preserve the integrity of both study types. The limitations mentioned in the sections covering single- and multi-generation studies concerning endpoints such as neonatal death, malformations, pre-implantation loss and resorptions also apply to developmental neurotoxicity studies.

Reproduction/Developmental Toxicity Screening Tests

42. In recent years, new screening tests for reproductive and developmental toxicity of shorter duration and using fewer resources have been developed. By definition, a screening test is limited in scope compared to a conventional test.

43. In the mid nineties, the OECD introduced Test Guidelines for screening tests for reproductive toxic effects. The "Reproduction/Developmental Toxicity Screening Test", (TG 421) and the "Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test" (TG 422) covers part of the Screening Information Data Sets (SIDS) for high production volume (HPV) chemicals. The TG 422 is a combination of a 28-days toxicity study and a reduced one-generation study whereas the TG 421 is a reduced one-generation study.

44. The purpose of the tests is to generate limited information concerning the effects of a test substance on male and female reproductive performance such as gonadal function, mating behaviour, conception, development of conceptus and parturition. It is not suggested as an alternative to, or as a replacement for, the existing Test Guidelines for single and multi-generation and developmental toxicity studies.

45. The dosing of the animals is initiated two weeks prior to mating and continues during mating. In the females dosing continues until the end of the study on postnatal day four. In males dosing continues for a total dosing period of at least 28 days. The number of animals per group is at least ten animals of each sex which is normally sufficient to provide at least eight pregnant females per group. Effects on fertility and birth are registered. Live pups are counted and sexed and litters weighed on day one and four post-partum. The parameters include, among others, a detailed histological examination on the ovaries, testes and epididymides of at least the highest dosed and the control animals

Other *In Vivo* Toxicity Tests

46. Other toxicity tests than those mentioned above could reveal effects that indicate a potential of a chemical to interfere with normal reproduction (e.g., TG 407, 408 and 409). Thus, in all the toxicological tests involving repeated dosing, including the carcinogenicity test, the gonads and accessory sex organs are subjected to pathological examination including histopathology.

47. The US National Toxicology Programme (NTP) has developed the reproductive assessment by continuous breeding protocol (RACB). In that protocol, pre-natally treated mice are continuously housed with untreated male mice and the cumulative number of offspring is measured by removing the female when noticeably pregnant and then returning the female to the males cage immediately after delivery to establish a pattern of repetitive breeding (Chapin *et al.,* 1997; Lamb (1985). This protocol allows the examination of repeated litters from each generation to monitor delayed effects on fertility and reproductive performance, which can be more severe with longer exposure of parental animals, or show small but biologically important decrements in successive litters in treated animals versus controls.

II. PRENATAL END POINTS

Overview of endpoints assessed

Laboratory Animal Studies

48. For the evaluation of hazard following exposure to chemical substances, a number of Test Guidelines are available that include peri-natal assessment of endpoints (growth retardation, structural abnormalities and lethality) following prenatal exposure. These include the prenatal developmental toxicity study (TG 414), the one-generation reproduction study (TG 415), the two-generation reproduction study (TG 416), the developmental neurotoxicity study (TG 426), and two reproductive/developmental toxicity screening tests (TG 421 and 422).

Critical Windows of Exposure

49. In the Test Guidelines described, *in utero* exposure to the offspring could occur during specified phases (*e.g.,* from implantation until one day prior to anticipated delivery, or during the major period of organogenesis and sexual differentiation, as in prenatal developmental toxicity studies) or throughout the entire gestation period (*e.g.,* from conception to delivery, as in single- or multi-generation reproduction studies or the reproductive/developmental toxicity screening tests). These broad testing scenarios are designed to screen for a variety of adverse outcomes of prenatal exposure by ensuring *in utero* exposure of the foetus across a number of critical windows of vulnerability. While it is possible in theory to identify a day or range of days during which prenatal exposure of test animals would result in a specific adverse developmental effect (Wilson, 1965; Selevan *et al.,* 2000), it is seldom that a Test Guideline could provide sufficient information to precisely identify the critical window of exposure that resulted in a particular adverse outcome. Although single versus repeated exposures may result in very different developmental outcomes, a conservative assumption is typically applied when these data are considered in hazard evaluation for risk assessment: that any observed pre-natal developmental effect could have resulted from a single exposure to the test substance at any time point during the exposure period.

Latent Effects

50. Latent (delayed) effects of developmental exposure are indirectly assessed in the developmental neurotoxicity study, since neurobehavioral and neuropathological assessments are conducted following a period of non-treatment (from approximately postnatal day 21 to 60). However, latent developmental effects are not assessed in the prenatal developmental toxicity study, since the foetal offspring are taken for examination prior to birth. In a multi-generation reproduction study, second (and subsequent) generations of animals are unique in that they have been exposed to the test substance throughout *in utero* and postnatal development, and assessed for toxicity through most of their early adult life, allowing for the potential expression of latent effects in adulthood (as limited in duration by study design). It is, however, extremely difficult to distinguish between latent adverse effects that result from prenatal exposure and those that have resulted from postnatal or adult exposures, or an accumulation of multiple exposures over the various life stages of the animal. A comparison of the responses between offspring of successive generations on the same study may provide some limited information to address this issue.

Methodological issues

Corpora Lutea

51. The maternal ovaries are removed and examined at the time of necropsy and *corpora lutea* are counted. The *corpus luteum* (CL) is a transitory endocrine organ formed from the thecal and granulosa cells of the postovulatory follicle. Progesterone is the major steroid produced by the CL and is necessary for ovum implantation and the maintenance of pregnancy (Nelson and Gibori, 1993). Information on the number of *copora lutea* is useful in the interpretation of data on the viability of conceptuses (see below, pre-implantation loss). *Corpora lutea* counts can be conducted with fresh or fixed tissue, and are generally performed with the aid of a dissecting microscope. Minimal dissection of the ovarian tissue is required. Counting accuracy can be affected by size differences among species (mice are somewhat challenging), but technical difficulties are not insurmountable. In prenatal developmental toxicity studies, maternal necropsies are conducted prior to delivery of the offspring; however, in reproduction studies, maternal necropsy does not occur until the offspring are weaned (postnatal day 21), resulting in further technical challenges in the accurate counting of *corpora lutea*. Additionally, multiple litters produced by a single dam will compromise *corpora lutea* count; in dams after multiple pregnancies, *corpora lutea* counts are not likely to be reliable, since *corpora lutea* remnants from previous pregnancies may be included in the count. Review of study data can provide indicators of possible inaccuracies in count for example when the number of implantations exceeds the number of *corpora lutea* reported within a litter; individual *corpora lutea* counts and derived implantation loss values from such a litter are generally not included in the statistical evaluation of group data.

Pre- and Post-Implantation Loss

52. Prenatal mortality may be grossly evident as a reduction in live litter size at the time of caesarean section or parturition. Additionally, for prenatal developmental toxicity studies, a detailed examination of uterine contents will reveal incidences of early and late resorptions, which are evidence of previous intrauterine deaths that have occurred after implantation. In prenatal developmental toxicity studies at caesarean section, very early resorptions may be observed and described as “empty implantation sites”. The uteri of dams that are necropsied at lactation day 21 in reproduction studies will appear similarly. For females that do not have visible evidence of implantation sites, further visual examination of the uterus is recommended, using techniques such as pressing the uterine tissue between two glass slides or staining with 10% (v/v) ammonium sulphide (Narotsky *et al*., 1997). Comparison of the number of implantation sites with the number of live and dead foetuses or neonates for each litter provides a means of quantifying post-implantation loss. Pre-implantation loss is quantified through a comparison of the numbers of corpora lutea and implantation sites for each dam. In a multi-generation study, increased pre-implantation loss could result from an adverse effect on gamete transport or function, resulting in failed fertilisation, or it can be a consequence of direct effects on the preimplantation embryo or indirect effects on the uterus or endocrine status of the dam. In a prenatal developmental toxicity test, where dosing is begun after breeding and fertilization, preimplantation loss is assumed to result only from the latter. Consideration of the exposure paradigm and the pattern of dose response may provide information that assists in the interpretation of data that indicate the presence of pre-implantation loss, but it is unlikely that the specific mechanism of toxicity can be identified without further study.

External Foetal and Neonatal Examinations

53. A visual external examination of foetal/neonatal animals is conducted at caesarean section or parturition, enabling the identification of gross defects in structure or activity, which may have resulted from *in utero* exposure to the test substance. Viability is assessed, sometimes with the use of lung flotation as a criterion to identify those neonates that died after birth. Individual body weight data, and sometimes crown-to-rump measurements, are recorded. Gender is generally determined at this time in rodents, by visual examination of the external genitalia and in some cases with measurement of anogenital distance, and may later be reconfirmed (*e.g.,* at soft tissue evaluation of foetuses). The integrity and closure of the palate can be confirmed by gently opening the mouth of each foetus and examining the interior, although this evaluation may be deferred until subsequent soft tissue examination.

Visceral Foetal Examinations

54. Typically, approximately half of the foetuses from each litter on the prenatal developmental toxicity study in rodents are killed and placed into a fixative that facilitates the firming of soft tissues and the decalcification of the skeletal structure. Examination of each foetus may be performed by serial sectioning procedures or by a combination of gross organ evaluation (either before or after fixation) and selected organ sectioning. Due to their relatively large size rabbit foetuses are not easily serial sectioned. For that reason, most laboratories will prefer to perform a detailed necropsy on each foetus to evaluate the structure and integrity of the thoracic and abdominal organs, with the addition of specialised procedures to examine the cranium (*e.g.,* using a coronal sectioning technique to visualise the internal structure of the brain, the eyes, and the nasal passages). This procedure may also be adapted for rodents, thereby providing information on soft tissue anomalies for 100% of the fetuses. Additionally, specialised procedures (*e.g.,* Staples, 1974), may be used, either as a standard method or to further explore possible developmental alterations in a known target organ. Histopathological evaluation of foetal tissues is not included in a standard guideline protocol, but could prove useful to further evaluate observed or anticipated anomalies.

Skeletal Examinations

55. For a prenatal developmental toxicity study in rodents, approximately one-half of the litter *(i.e.,* comprised of those foetuses that are not assigned to soft tissue examination) is typically processed for skeletal evaluation. For rabbits that are not serial sectioned (rabbit foetuses: 50% visceral examination with head; 50% visceral examination without head), all foetuses can be examined for skeletal abnormalities, after the soft tissue examination has been performed; rabbit foetuses must be eviscerated, skinned, and excess fat deposits (when present) removed from the shoulder area in order to facilitate proper processing. This procedure can also be adapted for rodents, thereby allowing skeletal evaluation for 100% of the foetuses. The foetuses (both rodent and non-rodent) are placed in a dehydrating fixative (*e.g.,* ethanol or acetone), macerated in potassium hydroxide solution, stained with calcium-specific Alizarin red S, and placed in a glycerine-based clearing solution to remove excess stain from soft tissues. The evaluation of foetal cartilage is recommended and can provide useful information on the cartilaginous precursors to ossified skeletal structures. Cartilage evaluation could entail double staining (involving an additional pre-clearing step in processing the foetuses, in which they are stained with cartilage-specific alcian blue dye) or alternative techniques, such as the use of differential Alizarin red S staining gradients. Double-staining techniques may be utilized in rodent studies but are not typically used in rabbit studies. Following

sufficient clearing of stain(s), the foetuses are examined for abnormalities of the cranial, thoracic, pelvic, and axial skeletal structures. For each foetus, the extent of ossification, normal structural formation (shape, size, integrity), position and articulation are evaluated.

Data interpretation

Malformations versus Variations

56. Classification of foetal and neonatal observations into malformations and variations is a common practice. A commonly used definition of a malformation is a permanent structural change, which may adversely affect survival, development, or function, while a common definition of a variation is a divergence beyond the usual range of structural constitution, which may not adversely affect survival or health (US EPA, 1991; Chahoud *et al.,* 1999; Solecki *et al.*, 2001, 2003). There is no generally accepted classification of malformations and variations, in part because there is a continuum of responses between normal and abnormal development. It is also noted that a given observation may be classified as a malformation in one species, and a variation in another, or the classification may change depending on the gestation day of examination. Additionally, some laboratories may attempt to further subdivide these definitions, e.g., by distinguishing between variations that are developmental delays and those that could be considered structural alterations. Therefore, the terms must be carefully and clearly defined by each individual laboratory and within the context of each study report.

International Harmonisation of Terminology

57. The International Federation of Teratology Societies (IFTS) Committee on International Harmonization of Nomenclature in Developmental Toxicology developed and published a glossary of internationally accepted common nomenclature to use when describing observations of foetal and neonatal morphology (Wise *et al.,* 1997). The purpose of this effort was to advance the harmonisation of terminology, and to reduce confusion and ambiguity in the description of developmental effects, particularly in submissions to regulatory agencies world-wide. Subsequently, the results of several terminology workshops, held in Berlin from 1998-2007, have been published in the literature (Chahoud *et al.*, 1999; Solecki *et al.*, 2001, 2003) and on the internet (http://www.DevTox.org). Familiarity with the internationally harmonized terminology for external, visceral, and skeletal observations, and appropriate use of the terminology in data collection, reporting, and review, is encouraged. It is recognised, however, that although the common nomenclature developed by this effort has been widely available and internationally accepted, there is no guarantee that the terminology has been uniformly used by all laboratories that conduct studies for chemical hazard assessment.

Interrelationship of Endpoints

• Maternal Toxicity

58. Studies intended to assess prenatal hazard are generally designed to include at least one dose group that elicits some degree of maternal toxicity. Endpoints of maternal toxicity, which are defined by study protocol, could include, for example, morbidity or mortality (which may be observed in spite of efforts to establish dose levels that do not result in these effects), altered gestation length, changes in clinical chemistry or haematology parameters, clinical observations, body weight, body weight change, food or water consumption, organ weights, gross necropsy, and/or histopathology data. There is a high degree of correlation between maternal condition and the status of the litter, which is particularly obvious

at very toxic dose levels. However, with few exceptions, it is not possible to fully distinguish between those effects on *in utero* development which are attributable to direct foetal exposure to the toxicant versus those effects which are due to, or exacerbated by, maternal toxicity. This is in part, due to the limited evaluation of maternal toxicity in these protocols (ECETOC, 2004). Further mechanistic studies are needed to elucidate the potential role of maternal toxicity as a cause of the developmental toxicity; some examples are provided by Daston (1994) and Hood and Miller (1997). Adverse effects on the developing organisms are, regardless of the cause, still toxic manifestations of treatment. For that reason, evidence of maternal toxicity does not automatically negate the observation of foetal toxicity at a similar dose level. Furthermore, two recent feed restriction studies in the rat and rabbit (Fleeman, 2005 and Cappon, 2005) clearly show that severe weight loss or decrease in body weight gain per see induced minor changes in skeleton development but no effects on viability or malformations in the rat. In rabbits abortions occurred in the most severe restricted group but no malformations. In some cases, however, the presence of maternal toxicity may impact the interpretation of study data. For example, when maternal toxicity is so severe (e.g., mortality) that foetal well-being is compromised, information on developmental effects may be difficult to interpret.

• Mortality and Incidence of Malformations

59. The prenatal developmental toxicity study design includes sacrifice of the rodent or rabbit dam one day prior to expected delivery, in order to ensure that malformed foetuses are not lost to maternal cannibalism (Schardein *et al.,* 1978), as could happen in a reproduction study. Nevertheless, even the prenatal developmental toxicity study does not allow the researcher to distinguish the source or cause of prenatal mortality. Intrauterine deaths may be the result of malformations that are incompatible with continuing viability. However, only those that occur in later stages of foetal development might potentially provide evidence of developmental anomalies (via the evaluation of late resorptions or dead foetuses that are not in advanced stages of maceration). While it may not be possible to make this distinction, the contribution of malformed foetuses to overall effect on litter viability can be appropriately analysed by combining the litter incidence of conceptuses that are malformed, resorbed (early and late), and dead (full term but nonviable at caesarean section) and performing appropriate statistical analyses of group values.

60. The sensitivity of the test for detection of rare events such as malformations is limited, due to the use of a relatively small number of animals. With the normal group sizes of 20 pregnant rats, it is not possible to identify any increase in major malformations unless high dose levels are administered or the substance studied is highly embryo/foetotoxic (Palmer 1981). To assess the developmental toxicity of a chemical, it is therefore important to include information on other developmental effects such as minor anomalies, variations, foetal death and growth. In addition, malformations of organs developing after the period of major organogenesis, *e.g.,* the sex organs and the brain, may not be detected if the study is performed according to the prior guideline for the teratology study. Examples of developmental toxicants that affect the morphology and function of the male reproductive system following exposure during the period of male sexual differentiation in late gestation include dibutylphthalate (Mylcreest *et al.*, 1999), linuron, a weak androgen antagonist (McIntyre *et al.*, 2002), and vinclozolin, an anti-androgenic herbicide (Gray *et al.*, 1999). Consequently, in the updated TG 414 the dosing period has been extended to cover at least the period from implantation to one day prior to the day of scheduled kill, which is one day before the expected day of delivery. (This extension of the treatment period, however, does not guarantee that the study will detect all changes that would indicate that a chemical is a potential endocrine disruptor, nor does it ensure that a negative outcome is evidence of no endocrine disruptive potential.). The dosing period should cover the entire period of gestation to the day of scheduled kill in order to also include the examination of effects from the pre-implantation period unless data on the chemical indicate an incidence of pre-implantation loss that would confound the evaluation of malformations.

• Litter Size and Mean Foetal Body Weight

61. The effect of litter size on individual and mean foetal body weight for that litter is well established. In polytocous animals, foetal or neonatal weights are generally inversely correlated with litter size, and the upper end of the dose- response curve may be affected by smaller litters and increased foetal or neonatal weight. While the average mean live litter size for untreated animals often falls within a narrow range, wide variations in litter size can occur spontaneously (MARTA and MTA, 1995, 1996). Trends towards larger litter sizes have been observed to occur over time, in part due to breeding programs that select for this trait. In addition to natural variation, litter size at caesarean section or parturition can be affected by prenatal chemical exposure, *e.g.,* through germ cell toxicity, implantation failure, or embryo-lethality. In the statistical analysis of mean foetal or neonatal body weight data, the litter weight should be statistically adjusted for the size of the litter, for example by using covariate analysis techniques.

• Factors that Could Influence the Incidence of Skeletal Variations

62. The extent of skeletal ossification at the time of death (*i.e.,* at caesarean section of the dam, which is scheduled for approximately one day prior to expected delivery) is determined for each foetus assigned to this evaluation. Skeletal development progresses along a standard and predictable time line (Spark and Dawson, 1928). Nevertheless, there can be normal variability in the schedule of ossification, *e.g.,* among various laboratory strains. For that reason, it is critical to establish a scientifically justified gestational day for caesarean section. Additionally, it is important to control, as much as possible, the time of day at which caesarean sections are performed, across control and treated groups, since the incidence of delayed ossification (generally classified as skeletal variations) can be directly related to the gestational age of the foetus and may not be an adverse effect of treatment. The epitome of scheduling would entail observation of the exact time of mating and performance of caesarean section at the precisely equivalent gestation duration (in hours) for each dam on study; this is, however, not practicable for most laboratories that conduct large guideline prenatal developmental toxicity studies. Nevertheless, it is important that the scheduling does not introduce a potential bias.

63. The extent of foetal ossification depends to some extent on foetal size, and it is a common observation that smaller foetuses (from larger litters) show an increased incidence of delayed ossification when compared with larger foetuses (from smaller litters) of the same dose group. In the statistical evaluation of ossification data, it may be appropriate to include adjustments for foetal weight or at least consider foetal weight when interpreting the data.

• Statistical Evaluations

64. Statistical analysis of fetal and neonatal data should be conducted with careful consideration of study design, the endpoint under consideration, sample size, effects of gender, and the influence of litter on analytical outcome. It is critical that littermates not be treated as independent observations in the statistical analysis; Holson and Pearce (1992) found that in the analysis of body weight data, treating as few as 2 litter mates per litter as independent observations can increase the nominal alpha by a factor of almost 3. For the selection of appropriate statistical methods and data analysis, consultation with a statistician familiar with appropriate analysis of litter data is recommended. While detailed discussion of statistical approaches is beyond the scope of this guidance document, specific issues of concern are raised here, and in the following chapter on postnatal endpoints.

65. The power of a study, that is, the probability that a study will demonstrate a true effect, is important in the evaluation of prenatal toxicity data. Factors that may influence the statistical power include the sample size used in the study (with the assumption that the litter is the basic unit of analysis), the background incidence of the finding, the variability in the incidence of the endpoint, the robustness of the data, and the method of analysis. Statistically, with a prenatal developmental toxicity study size that includes 20 litters per dose group, where all fetuses are evaluated, the minimum detectable change is:

- an increased incidence of malformations 5 to 12 times above control levels,
- an increase 3 to 6 times the *in utero* death rate, and a
- a decrease 0.15 to 0.25 times the foetal weight (US EPA, 1991).

For multigeneration studies, the detection of structural abnormalities in the F1 and F2 pups has been shown to be dependant not only on the number of litters assessed, but also on the number of pups from each litter that are examined for each endpoint, and on the degree of relatedness of the effects in one pup in the litter to another. Since the pups are not identical, there is statistical value (improved power) gained from examining all of the pups in a litter for a postnatal malformation, as is done in the developmental toxicity study. Examining many pups/litter in the F1 generation greatly enhances the ability to detect low dose effects. Even when litter mean values are analyzed, examining more than one pup per sex per litter can improve the statistical precision of the analysis (reducing the error mean square used to calculate the F statistic). In general, the size of "litter effect" is not the same for all endpoints in a multigeneration study, the size of the litter effect varies across dose (being larger at high, more effective dose levels), and the litter effect for an organ varies from one chemical mode of action to another.

66. Statistical significance does not need to be present to validate the biological significance of treatment-related effects. This is particularly true of findings with low incidence (i.e., rare malformations) or high variability, or in situations where the concurrent control data have an unusual incidence profile. In the same way, statistical significance does not necessarily signify biological significance, and scientific judgement and relevant historical control data should be used to distinguish between fortuitous and real findings.

• Concurrent/Historical Controls

67. Concurrent control data are required for every study. On the other hand, historical control data, which are generally comprised of well- characterised negative (vehicle) control data from multiple studies, are not required, but may nevertheless be available and considered useful and appropriate for interpreting study findings. Comparison of concurrent study control data with the data from treated animals should always take precedence over comparison with historical control data. If historical control data are used, the most appropriate of these are from studies conducted in the same laboratory, within a reasonable amount of time prior to the study being interpreted (*e.g.,* ± 2 years) in order to avoid genetic drift in the laboratory animal population, and under the same study conditions (*e.g.,* identical species, strain, source, age, vehicle, route and duration of administration, technical personnel, etc.). It is important that the data include sufficient information to render it meaningful in the context of the concurrent study. For example, definitions of terminology should be provided; incidental and continuous data should be fully characterised and summarised with appropriate data ranges, maximum, minimum, median, and mean values; data variance should be addressed. Historical control information that is compiled by animal suppliers or through surveys (or including genetic drift) across multiple laboratories (Clemens *et al.,* 1994, MARTA and MTA, 1995, 1996) can also be useful in some situations and under the appropriate caveats. Overall, the interpretation and use of historical control data requires careful consideration, and the application of scientific judgement and expertise. If historical control data are demonstrably different from concurrent control data, it may be an indication that the study contains some fatal flaw. In the most egregious case, it may not be appropriate to utilise the historical control data in the interpretation of data from treated groups.

III. POSTNATAL ENDPOINTS: NEONATAL GROWTH, DEVELOPMENTAL LANDMARKS AND FUNCTIONAL/BEHAVIOURAL NEUROTOXICITY

Overview of endpoints assessed

68. A number of Test Guidelines are available that include assessment of postnatal endpoints induced during development. These include the "One- and Two-Generation Study" (TG 415 and 416), the "Reproductive/Developmental Toxicity Screening Tests" (TG 421 and 422), and the "Developmental Neurotoxicity Study" (TG 426).

69. The endpoints assessed using the Test Guidelines are shown in table 2.

Table 2: endpoints assessed by the various TG

Endpoints	**TG 415**	**TG 416**	**TG 421 & 422**	**TG 426**
Birth weight	+	+	+	+
Survival-perinatal period	+	+	+	+
Survival-lactation period	+	+	-	+
Survival-adult	-	+	-	+
Growth-perinatal period	+	+	+	+
Growth-lactation period	+	+	-	+
Growth-adult	-	+	-	+
Physical development - sexual maturation	-	+	-	+
Functional development	-	(+)	-	+
Behaviour	-	(+)	-	+
Neuropathology	-	(+)	-	+
Reproductive functions	-	+	-	-

+ required; (+) optional

Methodological issues

Standardisation of Litter Size

70. Standardisation of litter size by random removal of pups a few days after birth to yield, as nearly as possible, a maximum of 8 pups (4 females, 4 males) by so-called "culling" is often used and the procedure is described in the Test Guidelines for generation studies and perinatal studies. Adjustment of litter size is optional in the TG 416, while the TG 426 recommends that litter size should be adjusted to a number close to the mean litter size for the animals used in the study.

71. An argument in favour of standardisation of litter size is that since pup weight is related to litter size it might lead to a more uniform pup weight at weaning. In a study of data from approximately 500 litters standardisation of litter size seemed only to increase the mean pup weight in the litters, while the variation remained unaffected (Palmer 1986). Also, culling may result in elimination of 25-40% of the offspring and may introduce bias for example by random elimination of runts (Palmer 1986).

72. The rate of growth and maturation of offspring may vary with litter size, but it is not clear whether variation in litter size also affects behaviour (Barlow and Sullivan 1975). Some studies have reported changes in behaviour in offspring from large litters compared to offspring from small litters, while others studies have found no differences (Lore and Avis 1970, all cited by Barlow and Sullivan 1975). A difference in litter size of 2-3 pups only was found not to be of any significance in any of these studies.

73. Litter size at birth plays a role for postnatal growth also in studies where the litters are standardized. Litter size should generally be considered in the statistical evaluation.

Route of Exposure

74. Study paradigms that include reproductive phases are generally subject to wide variations in exposure to the test subjects, as well as uncertainties regarding actual exposures to the offspring. Generally, even when more specific data are not available (*e.g.,* tissue levels of test chemical), it is assumed that an adverse treatment-related response in the parental or immature animals demonstrates adequacy of dose. Often, even in the absence of treatment-related effects in the offspring, it may be assumed that the offspring have been exposed to the test material during gestation (via maternal circulation) and lactation (via maternal milk). With most exposure routes, direct exposure of offspring to test substance will occur to some extent, for example, pups will begin to eat treated feed during their third week of the lactation period, and on a test-substance (mg)/body-weight (kg) basis they may actually be consuming a higher dose than the adults. When exposure is achieved via feed, pups of a single litter and housed together with their mother cannot be considered truly independent. This may have implications on the number of animals (F0) and offspring (F1) to be used. In inhalation studies that utilise chambers which expose the whole litter, pups receive inhaled and dermal doses simultaneously. When dosing via food or drinking water, the treatment is easily continued through the period of birth and in the neonatal period. As the pups gradually starts to consume food and water from around postnatal day 14, the exposure in the last part of the lactation period will be partly indirect via maternal milk and partly direct. Oral gavage or parenteral administration may have to be postponed for one or a few days around the period of birth.

75. When there is a need to ensure adequate exposure of the offspring during critical stages of development (*e.g.,* when assessing effects on the developing nervous or immune system) and/or to quantify exposure to pups, it may be necessary to consider the use of direct dosing of pups during some stages of the pre-weaning period**.** Direct dosing of pups, if done correctly by properly trained personal, does not produce stress in the pups. Careful consideration should be given to the impact of such procedures on toxic response and data interpretation (ILSI, 2003; Moser *et al.*, 2005).

76. Normally inhalation studies include exposure to both the dam and offspring. Separate inhalation exposure of the dam or pups only may be a possibility since indications that removal of the dam from the litter for six hours per day only causes a slight and reversible weight decrease during the first days (Pryce and Feldon, 2003). However, removal of the dam from the pups for six hours a day may induce overt physiological changes in the offspring, resembling those observed in animal models of maternal deprivation. Consequently, more studies illuminating the consequences and practical problems related to inhalation exposure of the dam only are needed before this possibility can be favoured. If used, it is recommended to include an additional control group, where the dams are not removed from the pups.

77. The dermal route of exposure is not recommended for reproductive toxicity studies. The technical difficulties associated with reproductive toxicity testing by administration by the dermal route outweigh the advantages of mirroring the normal human exposure. Other studies, such as ADME studies should be undertaken to facilitate extrapolation from the oral to the dermal route, if this is required.

Cross-Fostering

78. Postnatal effects of pre- and postnatal treatment may be the result of interference with the offspring prenatally, postnatally, or both. Developmental changes may not only be caused by the chemical affecting the developing organism before and after birth, but may also be induced via effects on the mother. For example, lactation or maternal care may be affected and potentially any alteration in maternal physiology or behaviour may affect the behaviour of the offspring. To control for this, cross-fostering techniques have been developed where the prenatally exposed litters are reared by non-exposed mother and vice versa. This obviously requires more animals and demands more resources, and is therefore not required in screening studies. However, it can be useful in follow up studies to further investigate effects observed in the initial screen.

Age of Pups

79. The age of the offspring may be calculated from the time of birth, *i.e.,* post partum age, or from the time of mating, *i.e.,* postcoital age. Studies have shown postcoital age to be a better predictor of age of appearance of developmental landmarks in the pre-weaning period than postnatal age, especially if the length of gestation periods differs among groups (Hughes 1986, Raimondo and Draghetti 1990). Postcoital age has been used in several studies of developmental neurotoxic effects (Goodlett *et al.,* 1987, Kelly *et al.,* 1988, Hass *et al.,* 1994a, b, Hass *et al.,* 1995). If the expected delivery day, *e.g.,* gestation day 22 for Wistar rats, is designated postnatal day 0 for all offspring (regardless of when they were actually born) the postcoital age is used, but the "age" also relates to the day of expected birth.

Time of Testing

80. The response of an animal in a behavioural test will depend on the time of day the test is carried out. Rats are nocturnal animals and waking of animals during their normal sleep period is likely to produce different behaviour than testing during normal wake period. Therefore, nocturnal animals may be tested at night or more conveniently reared under reverse lighting conditions and tested under red light during daytime (Barlow and Sullivan 1975).In any case, to get robust data, most important is to perform the tests under the same conditions.

Experiences of Offspring

81. The experience of offspring during infancy may affect their later behaviour. For example frequent handling of rats during infancy will alter their physiological response to stress (Levine *et al.,* 1967, Meaney *et al.,* 1988, 1991), and their behaviour in tests for emotionality and learning (Levine and Broadhurst 1963, Nunez *et al.,* 1995, Meaney *et al.,* 1988, 1991). Animals reared in an enriched environment may not only show behavioural changes but also have a heavier cerebral cortex, increased cholinesterase activity and altered levels of brain monoamines (Rosenzweig and Bennett 1969).

82. In order to control for environmental experiences during infancy, the conditions under which the offspring are reared should be standardised within experiments with respect to variables such as temperature, humidity, noise level, lighting, cages, handling and cage cleaning (Barlow and Sullivan 1975). Effects of experimental design changes, including enrichment of the environment, should be evaluated by comparisons to historical control data and when relevant by positive control studies.

Sex of Offspring

83. Sex differences in offspring behaviour are very common. It is therefore important that both sexes are tested and that the results from the two sexes are separated in the analysis of results from developmental neurotoxicity tests, otherwise effects may be masked.

Experimenters Influence on Results in Behavioural Tests

84. Studies have shown that the expectations of the experimenter in some cases may have a significant influence on the results obtained in behavioural studies (Rosenthal and Fode 1963). To completely rule out this influence, the experimenter should have no knowledge of which treatment group the animals belong to. The administration of this may cause some practical problems, for example when the experimenters are involved in analysing and evaluating the data during the study period. However, testing with no indication of treatment group is recommended to avoid experimenter bias.

Test Automation

85. Automation of behavioural tests may reduce experimenter bias and the need for manpower and increase the amount of information per experiment, allowing detailed data analysis for a better interpretation of results. Automated methods are required for the assessment of activity in the OECD TG 426 developmental neurotoxicity study. Commercial equipment is available for a number of behavioural tests. Automated data collection, however, may reduce the ability to categorize behaviour. For example, in an automated motor activity system using beam breaks as a measure of activity, "head bobbing" may be scored as locomotor activity if a single beam is repeatedly broken by the animals head movement. Generally, care should be taken to ensure that automation of behavioural testing takes into account behaviours which may affect the interpretation of the collected data.

Groups for Behavioural Assessment

86. The assessment of postnatal development and behaviour in the TG 426 includes the same groups as in the one- and two-generation study, *i.e.,* three treatment groups and a control group. However, in groups where clear-cut developmental toxicity effects (*e.g.,* markedly decreased litter size or increased postnatal death) are observed, behavioural testing may not be possible due to lack of pups or because the health of surviving pups is too compromised to perform behavioural testing. In cases of high maternal or pup toxicity, behavioural testing may not be necessary.

Physical Developmental Landmarks

87. The physical development of the offspring is normally monitored by recording body weight several times during the pre-weaning period and once or twice per month after weaning.

88. Other physical and functional parameters are followed by recording so-called developmental milestones or developmental landmarks. These observations often show "when" rather than "if", the various landmarks first appear and are used to assess delayed or accelerated developmental time courses for the specific parameters being studied (Lochry *et al.,* 1986). These tests evaluate the presence or absence of each parameter, usually over a period of successive days, beginning prior to, or approximately on the day of expected development.

89. Examples of frequently suggested physical developmental landmarks are ear unfolding, first coat, upper and lower incisor eruption, eye opening, full coat and onset of puberty. The reliability of observing these landmarks among different observers has been assessed in a study by Hughes and Palmer (1986). As a result of this evaluation, three physical parameters, *i.e.,* ear unfolding, upper incisor eruption and eye opening, were considered reliable and selected for use. As development of physical landmarks usually correlate to the body weight development, this should also be recorded at the same time as the physical landmark.

90. In the TG 416, *anogenital distance* (AGD) should be measured at postnatal day zero in F2 pups if triggered by alterations in F1 sex ratio or timing of sexual maturation. The AGD is longer in males than in females. Several studies have shown that hormones and suspected endocrine disrupting chemicals may change the AGD (*e.g.,* Gray *et al.,* 2001, McIntyre *et al.,* 2001). In addition, decreased AGD in male offspring has been shown to be predictive of other effects such as hypospadias and undescended testes later in life. Consequently, AGD measurement in the F1 generation, when chemicals with expected or suspected endocrine disrupting properties are tested should be considered. F2 pups are normally terminated at weaning, while examination of AGD in F1 newborns allows for a correlation with parameters recorded in the same animals at adulthood. If AGD is permanently changed (*i.e.,* at birth and adulthood) this would constitute a permanent structural change. AGD can for example be measured using a slide gauge or a stereomicroscope with measuring scale. It is import to establish clear criteria for the measurement for example from the centre of anus to the centre of the genital bud. In addition, the handling of the animals during the measure should be careful to avoid variation in the measure by for example stretching the region in some animals more than in others. Since the AGD may correlate with the body weight of the pup, a standardized approach for weight versus AGD should be considered when the AGD is used as a covariate in the statistical analysis.

91. Assessment of *nipple or areola retention* in male rat offspring is not included at present as a parameter in OECD Test Guidelines, however, the measure is sensitive to exposure to hormones and some hormonal disrupters (especially anti-androgens) **(**Gray *et al.,* 2001, McIntyre *et al*., 2001). In addition, nipple or areola retention has been shown to be predictive of *e.g.,* hypospadias and undescended testes. Consequently, it may be useful to include this endpoint in the F1 generation, when endocrine disruption properties are expected/suspected. The F2 pups are normally terminated at weaning, while examination of the F1 newborns allows for a correlation with parameters recorded in the same animals at adulthood. Permanent nipples in males constitute a permanent structural change, *i.e.,* a malformation. In female offspring 12 areolas are normally visible around postnatal day 13 while very few or none are visible in male offspring. As the development of fur in the animals makes it difficult, or impossible, to see the areolas it is important to establish the correct time for the assessment in the animals used for the study. Often only the presence or absence is registered in the males instead of the numbers per male. In most studies the frequency in control male rats is low, *e.g.,* below 5% and in such cases assessment of presence or absence may be of sufficient sensitivity. However, in some studies control values up to around 30% have been reported (Hellwig *et al.,* 2000) and in such cases assessment of presence only will be rather insensitive. It is recommended to use the number of areolas in each male for the assessment, especially when the control values are high.

92. Assessment of *sexual maturation* is included in TG 416 and TG 426. Assessment of the onset of puberty in females is done by inspection of vaginal opening. In rats, this occurs around postnatal day 30-35. Although the rat is the most common species employed in reproductive toxicity studies it is worth noting that

age at vaginal opening is not an index of puberty in the mouse, in this case investigators would have to use age at first estrus. In males, balano-preputial separation is normally used as indication of puberty. Observation of testicle descent relies on how the animal is handled during inspection and is rather difficult to assess. Balano-preputial separation corresponds to puberty in male rats (Korenbrot *et al.,* 1977) and is the endpoint included in the TG 426 and the TG 416. In the rat, this occurs around postnatal days 40-45. Both the age and the body weight of the animal at sexual maturation should be registered. Registration of sexual maturity in males and females need to be carefully evaluated. This requires some training in order to ensure that the animals are scored similarly each time. For that reason it is also preferable that the assessment is performed by one person (or a few persons) using the same criteria each time.

General Methodological Considerations for Conducting Neurobehavioural Measures

93. Developmental neurotoxicity testing (OECD TG 426) includes assessment of behavioural ontogeny, motor activity including habituation, motor and sensory function, and learning and memory. A number of Guidance Documents have been developed describing evaluation and interpretation of these neurobehavioural endpoints (OECD 2004; IPCS 2001).

94. The degree of qualitative and quantitative comparability between human and experimental data was assessed at a US workshop for lead, methyl-mercury, selected agents of abuse, phenytoin, PCB, ethanol, and ionizing radiation (Francis*,* 1990). For the qualitative comparison the following functions were evaluated: motor development and function, cognitive function, motivational/arousal behaviour, sensory function, and social behaviour. Although a number of limitations were identified with cross-species comparability, the degree of comparability was considered acceptable (Francis *et al.,* 1990). It was not considered possible to make definitive quantitative comparisons at this time, based upon the relationship of endpoints to dose, but there were indications that for some of the agents discussed, cognitive function appeared to be the most sensitive category. Dose-response data were often limited, especially for humans. Comparisons of administered effective doses revealed a wide range of differences across species (up to 10.000-fold difference), while comparisons using internal measurements of dose (*e.g*., blood or brain levels) showed a remarkable correlation (generally, a 1-2-fold difference). The findings at this workshop supported the assumption that, as for other endpoints of developmental toxicity, functional changes in animal studies indicate that an agent has the potential to alter development in humans.

95. The OECD Developmental Neurotoxicity Guideline (TG 426) provides flexibility in selection of the specific behavioural evaluations to be performed, and in many instances specifies only the functions to be assessed and the age at which assessments should be completed. For this reason, it is essential that each laboratory document the sensitivity of the procedures to detect treatment-related changes, and the specificity of the procedures to detect the function being evaluated. Testing procedures should be designed and implemented by individuals trained in the field of behavioural assessment, and technical staff should be trained in the conduct of the specific procedures used. Careful adherence to SOPs is essential for these assessments because changes in procedural details like animal handling, timing and methods of equipment cleaning, changes in the test environment (e.g., lighting or noise levels), etc can significantly affect the behavior of the test animals. Variability in test procedures can result in increased variability in test results which decreases sensitivity of the test to detect the effect of the test substance on behavior.

96. The experience of offspring especially during infancy may affect their later behaviour. For example, frequent handling of rats during infancy may alter the physiological response to stress and the behaviour in tests for emotionality and learning. In order to control for environmental experiences, the conditions under which the offspring are reared should be standardised within experiments with respect to variables such as noise level, handling and cage cleaning. The performance of the animals during the

behavioural testing may be influenced by *e.g.,* the time of day, and the stress level of the animals. Therefore, the most reliable data are obtained in studies where control and treated animals are tested alternately and environmental conditions are standardised.

97. Prior to implementation of a new behavioural test, the laboratory should evaluate performance of control animals to ensure a reasonable stability of baseline performance across test occasions and an acceptable level of variability in test performance. In addition, the laboratory should demonstrate its ability to detect treatment-related changes in behavior by conducting test with positive control chemicals for most procedures used including tests showing the laboratory's ability to detect both increases and decreases in behavioural performance. When changes in behavioural test procedures are implemented, historical and positive control evaluations should be repeated to ensure that the procedural changes have not decreased the sensitivity of the test to detect treatment-related changes.

98. Sample size for specific test parameters are recommended in OECD Developmental Neurotoxicity test Guideline 426. For most evaluations, a minimum of 10 animals/sex/dose group (one/sex/litter) is recommended (See OECD TG426 Developmental Neurotoxicity Guidelines 2007 and US EPA OPPTS 870.6300 Developmental Neurotoxicity Study, 1998). Sample size should be large enough to ensure sufficient statistical power to detect treatment-related effects for the parameter being evaluated, given the specific procedures used in the laboratory. Larger sample sizes may be needed if the behavior is poorly controlled or highly variable. For most behaviours, it should be possible to detect changes of 5-25% or less from control levels (Buelke-Sam *et al* 1985, CBTS pp 591-634). Test procedures or devices used to test adult animals may need to be adjusted for use when used to test neonates or young animals. For example, sensors (e.g. photocells for motor activity monitors) normally used to monitor adults should be adjusted to appropriate heights to detect movement when used for younger or smaller animals. In some cases, it may be necessary to adjust the size of the apparatus (e.g. the pool diameter in the Morris Water Maze) or the testing procedure (e.g. delay length in a memory task) to ensure similar test sensitivity between weanling or young rats and adult rats.

Measures of Behavioural Ontogeny

99. Behavioural ontogeny can be assessed by measuring the ontogeny of reflexes in developing animals (see examples in Table 3). Changes in the pattern of motor activity during the early postnatal period can also be used as an assessment of behavioural ontogeny. The measures should include two behaviours that do not develop at the same age. Where multiple measures of behavioural ontogeny are evaluated, each measure should be evaluated at multiple time points over the appropriate developmental stage in order to characterize the ontogeny of the behaviours.

100. Measuring functional endpoints requires training to ensure that all animals are scored using the same criteria. For this reason, it is preferable that each functional endpoint is scored for all animals by the same person or by a few persons that use the same criteria. If more than one observer is used to measure functional endpoints, training should be used to ensure that comparable scoring criteria are used and to ensure inter-observer reliability in the scoring criteria.

101. Because handling during the course of evaluation of physical or functional landmarks can influence the behaviour of the animals later in life, it is important that litters from all treatment groups are handled similarly. Handling animals only until a positive response is observed (e.g., eye opening) may result in differential handling of treatment groups (e.g., maximum handling of the most affected groups) and this handling may alter subsequent behavioural measures.

Table 3. Measures of reflex development/behavioural ontogeny

Test	Description	Endpoints	Test Age(1) (PND)	Reference
Surface righting reflex	Place pup in supine position on flat surface. Measure time to right or yes/no within fixed time	Neuromotor abilities	2-4	Altman & Sudarshan, 1975
Negative geotaxis reflex	Place pup head downward on inclined plane. Measure time to turn 180° or yes/no within fixed time	Reflex development Neuromotor abilities	7-10	Adams, 1986
Homing reflex	Place pup between home bedding and clean bedding. Measure bedding choice and time to choose	Reflex development Neuromotor abilities, Sensory function (odour)	6-10	Adams, 1986
Air righting reflex	Drop animals from supine position. Measure ability to land on four feet, yes or no	Reflex development Neuromotor abilities	12-17	Altman & Sudarshan, 1975; Adams, 1986
Auditory startle reflex	Present sudden noise to pup. Measure startle response, yes or no	Reflex development Neuromotor abilities	10-14	Buelke-Sam & Kimmel, 1979
Swimming ontogeny	Score different stages of swim development. Measure body position, limb usage, direction of locomotion	Reflex development Neuromotor abilities	6- 25	Schapiro *et al.,* 1970; Vorhees, 1983
Motor activity	Place pups in activity chambers. Measure spontaneous activity and within-session habituation	Neuromotor abilities	13-21 (usually 13, 17 and 21)	Ruppert *et al.*, 1984, 1985

(1) Note that the test ages are general recommendations and that individual laboratories should establish the relevant ages for each test

Measures of Motor Activity

102. Motor activity tests provide measures of overall motor activity as well as measures of non-associative learning (habituation). Motor activity can be evaluated using a range of tests. Table 4 provides some examples of the types of test that have been employed. OECD TG426 recommends that activity tests be conducted in automated equipment and that animals should be placed individually in the device for a defined period of time. OECD 426 also recommends that motor activity should be evaluated in at least pre-weaning

and adult stages of life. Habituation in motor activity is defined as a decrease in activity over time within a single test session, and can be evaluated by comparing activity levels in consecutive intervals of the test session. Motor activity scores should reach asymptotic levels by the end of the test session. Motor activity testing in the pre-weaning animals may also be used to evaluate behavioural ontogeny since normal intra-session habituation of motor activity is usually absent in very young animals (e.g., post natal day (PND) 12) and develops in weanling animals (see Table 3). Tests sessions should be of sufficient duration to detect normal intra-session habituation of motor activity. This generally means that short terms tests (e.g., 2-10 min) should not be used because habituation will not normally occur in these short duration tests.

103. The *open field* and the *hole-board* have generally been used to measure short-term activity and exploration of new environment (generally between 2 and 5 min test session). The methods may, however, be adapted for testing of longer-term activity and habituation and would in such cases be useful for testing according to the TG 426.

104. Automated devices such as the figure-8 maze, radial arm maze, and new cages similar to home cages can be used for measuring activity over longer time periods as well as habituation. Within- and between-laboratory reliability have been demonstrated for a variety of devices across several different types of chemicals (Crofton *et al*., 1991, Moser *et al*. 1997, Buelke-Sam, 1985).

105. Open Field: The open field usually consists of a rectangular or circular arena. The animal is placed into the arena for a defined period of time and measures of motor activity are recorded. Examples include ambulatory counts or distance, rearing (two front paws simultaneously off the floor) or vertical counts or time spent in rearing. Measures of motor activity patterns such as forays, as well as location of motor activity (periphery or central area of open field). Number of rears may also be counted. Activity levels in the open field can be evaluated for short periods (e.g. 2 minutes) that are appropriate for use in detailed clinical evaluations or longer trials (e.g., 30 min) as an evaluation of motor activity and habituation. Testing of activity in a novel open field can be viewed as a special type of open field test. This involved the same testing except the test chamber is novel for animals. For animals never tested in open field activity tests, the open field tests will in fact be novel chamber tests. As with other activity tests, animals are placed in the chamber and activity is monitored by photocells or video based recording systems. Activity can be quantified in a variety of ways, most frequently as the number of movements or time spent in movement during a specified interval.

106. Figure-8 maze: The Figure 8 maze can be viewed as a variant of the open filed. Due to the nature of its design, the Figure 8 maze is thought to provide a measure of motor activity less influenced by environmental factors. This device consists of a set of alleys arranged as a figure-8, usually with two additional dead-end alleys projecting outward from the center crossing. Movement of animals in the maze is recorded by photobeams located in the maze, and activity is evaluated as the number of beam-breaks during a specified period. Video-based recording system can also be used to measure movement of animals within the figure 8 maze. Patterns of activity can also be evaluated by tabulating non-consecutive beam breaks (often defined as locomotor activity) or the proportion of activity in various locations of the maze (e.g., the proportion of activity in the dead-end alleys). The Figure 8 maze has also been shown to be sensitive to treatments (e.g., d-amphetamine, chlorpromazine, methylmercury) that produce increases or decreases in activity (Reiter and MacPhail 1982; Elsner *et al*, 1988; MacPhail, 1999).

107. Hole Board: The hole board is used to measure both motor activity and exploratory patterns. The hole board is similar to the open field except the floor has holes that are sufficiently large for animals to insert their nose but do not permit animals to enter the holes. Various configurations of the hole board may be used. Exploratory activity is quantified, as time spent investigating the holes. By placing food below some of the holes in the hole board, the hole board task can also be used as a memory test comparable to the radial arm maze (File, 2001; Van der Staay, 1999; Brosnan-Watters, 1997).

108. Radial Arm Maze: Although developed to assess spatial memory, the radial arm maze can also be used to assess motor activity (Olton, 1987). Procedures for the radial arm are described in the section on tests of learning and memory below.

Table 4. Measures of motor activity

Test	Description	Endpoints	References
Open field	Square or circular box. Locomotion, rearing	Short or long-term activity Exploration	Adams, 1986; Denenberg, 1969; Schiorring, 1979; Walsh, 1976
Hole board	Box with holes in floor. Sometimes objects underneath.	Short or long-term activity Exploration	Adams, 1986
Radial arm maze	Eight arms extending radially from central hub	Short or long-term activity Locomotion, Learning	Adams, 1986; Walsh & Chrobak, 1987
Figure-8 maze	Several interconnected alleyways - forms a figure 8	Long-term activity	Adams, 1986; Reiter & MacPhail, 1982; Crofton *et al.,* 1993
Novel cage	Standard cages (non-home) within series of infrared beams	Long term activity	Gårdlund *et al.,* 1991

Measures of Motor and Sensory Function

109. Motor and sensory function should be evaluated least once during- the adolescent period and once during the young adult period (see OECD TG46). These assessments should be quantitative.

110. Until recently, measures of *sensory function*s in developmental toxicity studies have been limited to categorical assessments (presence or absence of response) rather than more sensitive measures of the magnitude of the response. However, more sophisticated automated behavioural techniques have become available that permit more sensitive quantitative assessment of sensory functions (table 5). These quantitative test methods are recommended over qualitative tests as they provide more sensitive measures of sensory functions.

111. Tests of neuromotor ability are often used in the evaluation of the ontogeny of reflexes or co-ordinated movements. Measures of reflex and motor development are the most widespread of all functional endpoints assessed in behavioural teratology studies (Adams, 1986; Buelke-Sam and Kimmel, 1979). The procedures for most of these commonly used measurements have been described in several reviews (Barlow and Sullivan, 1975; Adams, 1986); therefore only a brief description is given in table 3. Rotarod (and the accelerod as a variant), grip strength, beam test, and hind limb splay can be used for assessing neuromotor abilities (table 5).

Table 5. Measures of neuromotor and sensory abilities(1)

Test	Description	Endpoints	Age at testing	References
Rotarod and Accelerod	Placed on rotating rod at fixed velocity. Measure time on rod or yes/no to fixed period of time. Accelerod uses same approach but rotating rod increases speed. Accelerod tends to be more demanding and sensitive test.	Neuromotor abilities, Vestibular system, Motor coordination	Usually after weaning but may be conducted before weaning	Jones and Roberts, 1968; Kaplan and Murphy, 1972; Bogo *et al.,* 1981; Adams 1986; Pryor *et al,*1983a; Bushnell *et al.,* 1994
Grip strength	Duration of hanging on thin wire or special apparatus. Fore- and hind-limb grip strength	Neuromotor abilities	PND 13 or later	Meyer, *et al.*, 1979; Pryor *et al.,* 1983a; Crofton *et al.,* 1990, 1994a, 1994b
Hind-limb splay	Dropped from 30 cm above a sheet of paper. Distance between prints of hind paws	Neuromotor abilities	Usually after weaning	Gabriel *et al.,* 1991
Beam test	Animals placed on narrow beam and must traverse to reach escape platform or home cage. Measure time to traverse, slips, falls.	Neuromotor abilities	Usually after weaning	
Negative geotaxis reflex	Animal placed face down on 45 degrees surface, measure ability and time to re-orient with face upward	Neuromotor ability, vestibular function	PND 7-10	Pryor *et al.,* 1983a
Prepulse modification of startle response	Hearing ability Vision tactile	Quantitative measure, reflex modification, sensorimotor integration	Measure near weaning or after	Adams *et al.,* 1986;Crofton *et al.*, 1990
Auditory startle	Auditory sensory reflex response	Quantitative or qualitative	After PND 10	
Auditory startle habituation	Loud tone presented at regular intervals. Amplitude of startle response.	Habituation (non-associative learning)		
Eye-blink conditioning	Present cue paired with air puff that produces reflexive eye blink	Hearing ability Vision, Tactile, learning	Usually after weaning	Stanton & Freeman 1994
Operant techniques	Range of procedures using trained response to determine sensory thresholds or sensitivity (e.g., conditioned avoidance,	Can be used to assess wide range of sensory functions	Usually after weaning and usually in adults where food motivation used	Elsner, 1991; Gabriel *et al.,* 1991; Bushnell *et al.,* 1994; Pryor *et al.,* 1983b

(1) Note that some measure of reflex ontogeny are also measure of sensory or neuromotor function.

112. The negative geotaxis test is used to evaluate vestibular function. It is conducted by placing a rat nose downward on an inclined plane (usually about 60°) and measuring the time for the rat to rotate 180° to a position of nose upward. Although this test works well for assessing sensorimotor function in young animals, it appears to have limited usefulness in tests of adults because of large variability (Pryor *et al.,* 1983a).

113. The rotarod test can be used to assess balance and motor coordination. The rotarod test (Bogo *et al.,* 1981) is conducted by placing rats on a moving, motor-driven cylinder and measuring the amount of time the rat can maintain balance on the rotating cylinder. The test can be conducted using either an accelerating cylinder (accelerod) or a fixed speed cylinder. For the accelerating cylinder procedure, the speed at which the rat falls into bedding is measured while for the fixed speed procedure, the time to fall from the cylinder or performance after a fixed time period is recorded (e.g., 60 sec). The accelerating version of the rotarod test is thought to be more sensitive than the non-accelerating version of the test.

114. Quantitative behavioural measurements of sensory function are obtained primarily through reflex modification tests or discrimination tests of conditioned behaviours. By teaching animals to perform specific responses to sensory cues, the ability to detect sensory cues can be evaluated. By using olfactory, auditory, or visual cues it is possible to determine if animals can perceive the cues or changes in cues (e.g., variations in sound frequency). When the response is unchanged after stimulus parameters are changed (e.g., sound or light frequency or intensity) it indicates that the animals was unable to detect the change in stimulus properties. By varying the quality or intensity of the stimulus, it is possible to measure thresholds of detection for sensory function.

115. *Auditory startle habituation* involves repetitive elicitation of the startle reflex. With continued repetition at regular intervals, the magnitude of the startle response decreases indicating habituation to the startle-eliciting stimulus. This task can be used to evaluate both the reflex performance as well as simple non-associative learning (i.e., habituation). In rodents the procedure is usually conducted using about 50 startle trials with response magnitude averaged into blocks of 10 trials/block, and changes in response amplitude evaluated over the course of the test session. Parameters such as loudness and duration of the startle stimulus, inter-trial interval, and background noise levels should be optimized to maximize the sensitivity of the procedure for the laboratory, strain, and age of animal being tested. Startle responses have been demonstrated in most mammalian species. Note that although auditory startle is the most frequently used startle response, other reflexive startle response may be also used but the procedures should be and sensitivity of the procedures should be well documented and validated.

116. *Pre-pulse inhibition of startle response* can be used to evaluate the integration of sensory information. (Ison, 1984; Crofton and Sheets 1989; Koch, 1989) Pre-pulse inhibition is an inhibition of startle response magnitude when the startle-eliciting stimulus is preceded by another stimulus that does not normally elicit a startle response. Pre-pulse inhibition has been shown in most mammalian species. Typically the pre-pulse stimulus itself should not produce a startle response. Pre-pulse stimuli are normally presented 30-100 ms before the startle stimulus and the magnitude of the startle response is normally decreased in an intensity-dependent manner. In addition, the startle response latency is normally unchanged or increased after presentation of pre-pulse. Pre-pulse inhibition is thought to reflect the integration of multiple sensory inputs and disturbances in pre-pulse inhibition reflect disturbances in the ability to integrate information (Koch, 1989, Ison 1984). By modifying properties of the pre-pulse stimulus, it is possible to assess changes in sensory capacity to detect in the pre-pulse stimulus (Crofton *et al.,* 2000; Wecker *et al*., 1985).

117. Eye-blink conditioning is a form of classical or Pavlovian learning and involves presenting a conditioned stimulus (typically a tone) immediately before an unconditioned stimulus (typically a brief air puff to the eye). The conditioned stimulus normally does not elicit any response by itself. The unconditioned stimulus normally elicits an unconditioned response (e.g., eye-blink response); however after repeated presentation of the conditioned stimulus (e.g., tone) immediately before presentation of the unconditioned

stimulus (air puff), the conditioned stimulus (tone) will elicit the eye-blink response in the absence of the unconditioned stimulus. This learned eye-blink response is a conditioned response and provides a measure of basic learning capacity. (see Stanton and Freeman, 1994 for review). Methods for studying eye-blink conditioning exist in rats, rabbits and humans. Comparisons of studies involving humans and animal models indicate that the effects of a number of biological variables are similar across species and available evidence suggests that the biological mechanisms are similar across species. Although the eye-blink conditioning model has mainly been used for studying learning effects, it is possible to use this test to evaluate sensory system function by modifying properties of the conditioned stimulus (e.g., frequency, intensity, etc) and assessing its effect on the conditioned response.

118. In addition to the procedures described above, operant conditioning procedures can be employed to evaluate sensory function. Though not completely independent, classical conditioning involves presentation of the conditioned stimulus before presentation of the response while operant conditioning involves presentation of the unconditioned stimulus (e.g., food reward) after the response (e.g., bar press). Similar to the eye blink response procedure, operant test procedures used to assess sensory function are based on evaluating the animal's capacity to detect changes in properties of sensory cues (e.g., tone frequency or intensity, light frequency, spectral properties, etc) that control responses. Because operant procedures permit considerable flexibility in the use of a wide range of sensory cues to control behaviours, operant conditioning provides some of the most sensitive tests available to assess sensory function.

119. *Conditioned avoidance response* (CAR) involves training animals to avoid an aversive stimulus that is predicted by a warning signal (e.g., tone or light). The warning cue is presented before the aversive stimulus and animals must learn to avoid the aversive stimulus by making a response. Animals can be easily trained to perform a wide range of behaviours (crossing a wall, pressing a bar, climbing a pole) to avoid the aversive stimulus. For example, rats will learn to climb or pull a pole to avoid a foot shock, which is preceded by a warning signal (the conditioned stimuli) (Pryor *et al.,* 1983a, b). Since the conditioned stimuli can be a wide range of stimuli (visual, auditory, olfactory), changes in the intensity or frequency of the warning cue can be used to evaluate the ability of animals to detect or discriminate changes in the sensory warning stimulus. For example, using pole climbing as the measured response and auditory stimulus as the cue, it has been shown that toluene-exposed rats had a hearing deficit of frequencies above 8 kHz. Experiments with positive control substances have indicated that the CAR procedure is sensitive and capable of detecting specific sensory loss.

120. Operant conditioning techniques can also be used to evaluate somatosensory function. For example, Elsner (1998, 1991) developed a tactile-kinaesthetic performance test for rats by training rats to press a force-sensitive lever. By decreasing the range of force on the lever that earned reinforcers over test trials, this procedure can be used to assess the capacity of rats to control the force applied to the lever as a measure of somatosensory function. This procedure was successfully used to demonstrate that methyl mercury affected the tactile-kinaesthetic system of rats -including a disturbance in fine motor control (Elsner *et al.,* 1991).

Measures of Learning and Memory

121. Associative learning and memory should be tested twice, with one test conducted in post weaning offspring with one test conducted in adults. While there is considerable flexibility in the choice of specific tests to use, the tests must a) demonstrate learning via a change in behaviour either across tests sessions or within a single test session and, b) the test should incorporate measures of both learning (acquisition) and (retention) memory. While it is permissible to use the same animals at both times, it is generally recommended that different animals be used for each test to minimize the potential that learning in one tests confounds measures

of learning in the second test. If the same test is used in young and adult animals, it is unlikely that learning can be adequately assessed at both test times using a single group of animals.

122. A large number of associative learning tasks are available to assess learning and memory. Generally they can be classified into either appetitively or aversively-motivated tasks (Lochry *et al.,* 1986). Appetitively-motivated learning employs the use of reinforcers to maintain responding (i.e., animals exhibit some response to gain a reinforcer). Typically food or water deprivation is required to control levels of motivation in these tasks although preferred substances like sucrose or saccharin are also used to motivate animals. Aversively-motivated learning employs the use of aversive, unpleasant or painful stimuli to motivate responding (e.g., footshock, bright lights, loud noise, water). These aversive tasks can be further divided into either escape or avoidance tasks, depending on whether animals have the option to completely avoid the aversive stimulus by responding before its presentation or whether animals can only escape aversive stimulus. While appetitively motivated tasks normally require either food or water deprivation, aversively motivated tasks (avoidance tasks, water mazes) do not require such deprivation. Because of the potential impact of food or water deprivation in young developing animals, appetitively-motivated tasks requiring deprivation are not normally conducted on immature animals.

123. Some tests that would satisfy the criteria for associative learning are shown in Table 6: active or passive avoidance, taste or odour aversions, olfactory conditioning, various water maze tests including Morris Water Maze, Biel or Cincinnati maze, radial arm maze, T-maze, as well as a variety of operant-based learning and memory tests (e.g., delayed-matching-to-position, delayed spatial alternation, alternation tasks, delay tasks, etc). Additional tests are described in the literature for weanling and adult rats.

Table 6. Measures of learning and memory abilities, examples (see text for more details)

Test	Description	Potential Endpoints (1)	Reference
Passive avoidance	Animal must learn to withhold responding to avoid aversive stimulus	Trials to criterion, errors, latency to respond, latency to enter dark compartment	Adams 1986; Wier *et al.,* 1989; Lochry and Riley 1980
Active avoidance	Animals must perform a response (e.g., cross door, press bar) to avoid aversive stimuli.	No. of trials to criterion, No. of successful avoidances, No. of escapes where avoidance failure	Adams 1986
Operant conditioning (Skinner box)	Animals must perform a response to obtain reinforcement or avoid aversive stimulus. Wide range of test schedules, stimuli and response can be employed.	No. of correct responses, No reinforcers, percent correct choices, errors rates, (depends on specific test)	Adams 1986; Hass *et al.,* 1994a
Radial arm maze	Eight arms extending radially from central hub. Reinforcement in each arm.	No. of errors, i.e. revisiting an arm, activity, Trials to criterion. time	Adams 1986
Biel water maze, Cincinnati water maze	Maze with 6 sequential choices between left or right turn	No. of errors, latency to goal	Peele *et al.*, 1990
Morris water maze	Find platform hidden under the surface in large circular pool using spatial cues. Important to use different start positions.	Swim length and path, latency to escape platform	Kelly *et al.,* 1988; Hass *et al.,* 1995
T-maze	T-formed maze, select one arm of two with reinforcement. Can be dry or water based T-Maze.	No. of errors, latency, trials to criterion	Wier *et al.,* 1989; Lochry and Riley 1980

(1) Note that many of these tests can be used to asses a wide range of endpoints, depending on to goal and specific use of the tests. The endpoints listed here are to provide an overview of some of the more common endpoints that can be evaluated with the tests

124. Passive avoidance tasks require animals to withhold a response in order to avoid the presentation of an aversive stimulus (unlike active avoidance tasks where animals must make some active response). For example, one frequently used passive avoidance task involves avoidance of one side of a two-compartment chamber that has been previously associated with an aversive stimulus such as footshock. The avoidance of the area associated with footshock measured by latency to enter the shock–associated chamber, provides a

measure of associative learning. Passive avoidance tasks are technically easy and rapid to conduct, often with few learning trials required. However, they tend to produce variable results and may be sensitive to impairments in motor function. The sensitivity of this test may be increased by using repeated trials passive avoidance, where the endpoint is for example the number of trials needed before the animals do not enter the dark compartment (Robbins, 1977). Active avoidance tests are comparable to passive avoidance tests except that animals must make an active response to avoid an aversive stimulus. Typically a cue is associated with an aversive stimulus that predicts the onset of the aversive stimulus and the animal makes a response (move to another location, press a bar) to avoid the aversive stimulus. Like passive avoidance tasks, active avoidance tasks are learned rapidly but may also be sensitive to motor disturbances. Learning can be measured by the number of avoidance responses or the latency to respond.

Water mazes

125. A number of water maze tasks have been developed to evaluate learning and memory. All water mazes rely on escape from water as the primary motivation and are therefore aversively-motivated tasks. Many of these are swimming versions of previously developed appetitive mazes (e.g., T-maze), while others have been developed to measure specific learning and memory capacities (e.g., Morris water maze) (see Vorhees, 1987; Vorhees *et al.,* 1991; Morris, 1984; Brandais *et al.*, 1989; D'Hooge *et al.*, 1991 for reviews). Water mazes have a number of advantages that including no requirement for food deprivation, can be used in young animals, are rapidly learned, are free of appetitive or taste confounds, relatively insensitive to odour cues, and, within limits, also relatively unaffected by differences in body mass (Vorhees *et al.,* 1991).Because water mazes can be motorically demanding, it is important to consider potential motor disturbances especially in water mazes that rely on latency or speed measures rather than choice accuracy measures.

126. The *Morris water maze* was developed to measure spatial learning and memory in rodents (Morris, 1984; Brandais, 1989, D'Hooge *et al.,* 2001). This maze consists of a large circular water tank with an escape platform hidden below the surface of the water. While a number of test variations have been employed the most typical approach is to place animals into the water maze from one of 4 random locations and animals must locate the hidden escape platform to escape from the water. Since start position changes on each test trial animals must learn to locate the hidden escape platform based on extra-maze spatial cues. (Morris 1981, Morris 1984). To ensure that animals use spatial cues to locate the platform it is critical that the starting positions are changed between trials. Measures of learning are swim latency to locate the platform, swim distance to locate the platform and swim speed. The decrease in swim distance over trials generally provides a better measure of learning than swim latency, as the latter may be influenced by the swim speed of the animal. An advantage of this test is that it can provide information on the swim speed (motor capacity) of the animal directly during the learning trials. The swim path may also be recorded and used for a more detailed analysis of the performance of the animal. Measures of memory are normally conducted on the last day of testing in which the escape platform is removed and the time spent in the previous location of the platform as a measure of memory. The task has been shown to be sensitive to a variety of experimental insults to relevant brain structures and to age related impairments. It has been used to demonstrate effects of for example prenatal or early postnatal exposure to ethanol, technical xylene, N-methylpyrrolidone, toluene and malnutrition in rodents (Blanchard *et al.,* 1987, Kelly *et al.,* 1988; Hass *et al.,* 1994a, b, Hass *et al.,* 1995). Other variants of the Morris water maze can be used to evaluate working (or short-term memory) by conducting two trial per day with the platform location changing from day to day but not between the two daily trials. Thus the first trial provides information on the platform location and the second trial measure memory of that information. Measures include latency to locate the platform, swim distance and swim speed. Another version of this task allows the use of local cues, which may be used to test cued learning, or as a control task to evaluate motor function.

127. The *Biel water maze* consists of six sequential left-right choices from the start to the goal of the water maze. Testing in the Biel water maze may consist of initial trials in a straight water channel to assess swimming ability followed by testing in the six sequence Biel maze, followed by additional trials in which the position of the start and goal box is reversed from learning trials Normally, the measures used to asses learning and memory are the number of errors on each trial, i.e. turning left instead of right and vice versa. While maze times are normally also recorded for each trial, this measure is less valuable for evaluating learning than measure of choice accuracy. The Biel water maze has been shown to possess good reliability at reproducing effects from various chemicals, to be sensitive for detecting effects with a coefficient of detection in the order of 10-20%, and to be valid for reflecting the effects of neurotoxic agents.

128. The apparatus used for *T-maze* testing consist of a start box, left and right maze arm, and a choice point connecting the start box to both maze arms. The T-maze may be used either as a dry maze, where the animals are given a reward for choosing the correct arm (appetitively motivated), or as a water maze (aversively motivated). The procedure often used for testing is position habit task, i.e. the same arm is the correct choice throughout all trials. Other possibilities include delayed alternation, where the correct arm alternates from left to right, and discrete trials delayed alternation, where each trial consist of two runs. The first run is forced, i.e. the animal can only choose one side of the maze, while the second run is a free-choice with the arm alternate to that entered on the forced run as the rewarded choice. Delays between trials can be varied, to evaluate memory over short or long time periods; longer delays can also be used to distinguish performance from memory deficits. Learning that the same arm is always the correct one may be relatively simple for rats, and studies have shown that this task may not be affected by chemical exposure, while results of T-maze delayed alternation testing clearly indicate impairment of cognitive processes (Robbins, 1977).

129. Spatial learning processes have been strongly linked, in the rat, to the hippocampus and related structures (Olton *et al.,* 1979, Morris *et al.,* 1982). Two commonly used methods for assessing spatial learning are the Morris water maze and the radial arm maze.

130. The *radial arm maze* is an appetitively-motivated task and the apparatus consists of eight arms extending radially from a central hub. Normally, a reinforcement is placed at the end of each the eight arms and test trials consist of placing animals in the center and allowing them to locate the reinforcers at the end of each arm of the maze. The radial arm maze permits the assessment of both short-term (working memory) and long term memory (reference memory). Measures of memory include the number of errors (revisits to arms already visited) The radial arm maze have been shown to be sensitive for evaluating the effects of chemicals such as ethanol and trimethyltin (Miller *et al.,* 1982), of maternal phenylketonuria and prenatal phenytoin (Weisenburger *et al.,* 1990), and of methylnitrosurea-induced microcephaly (Akaike *et al.,* 1988).

131. Schedule controlled operant conditioning offers a wide range of test conditions to evaluate learning and memory, as well as sensory functions as described earlier. Operant test have been developed to assess other cognitive functions such as attention, response inhibition (impulsivity), discrimination, spatial and non-spatial memory. Because the range of operant tests available permits considerable control over learning and response parameters, these tasks can provide highly sensitive tests of specific cognitive functions. However, because operant tests require tight schedule control over behavior, training for these tasks may be excessively time consuming for initial evaluation of learning and memory in a developmental stud*y*. Although operant tests are typically conducted using a*ppetiti*ve reinforcers, aversive stimuli can also be used in operant test situations. The more frequently used appetitive versions of operant tests also require food or water deprivation and, combined with the extensive training usually required, are generally not suitable as tests of learning and memory in young animals. While far from a complete list, some examples of operant tests are provided below. (General refs: Buccufusco 2002; Laties, 1978)

132. *Spatial discrimination learning*: animals must learn to select a response (usually choose one of two

levers) based on spatial position in the operant chamber (e.g., right or left choice) in order to earn a reinforcer.

133. *Cued discrimination learning*: animals must learn to select a response (usually choose a lever) based on cues (tone, light) with the cue determining the lever that will earn a reinforcer. Once the discrimination is learned, cues can be modified (e.g., tone frequency or light frequency) to evaluate if animals can detect changes in sensory properties. Usually, the number of errors, trials to a fixed criterion and the type of errors is measured.

134. *Alternation tasks:* animals must learn to alternate responses from one trial to the next trial. Alternation may be spatially based (e.g., right versus left) or cure based (illuminated lever versus non-illuminated lever). Because the correct choice on one trial is dependent on the correct choice on the previous trial, animals must remember the choice made on the previous trial. By increasing the interval between trials, this task permits the measurement of memory. Usually, the number of errors, trials to a fixed criterion, the type of errors is measured, and the delay at which criterion responding is no longer maintained is measured.

135. *Delayed spatial alternation*: this is a special version of an alternation task in which choice is based on spatial location (right versus left) and delays are introduced between trials to assess memory. Usually, the number of errors, trials to a fixed criterion, the type of errors, and the delay at which criterion responding is no longer maintained are measured.

136. *Matching to sample and non-matching to sample tasks*: animals must learn to respond to a lever based on a cue delivered immediately before the choice. In matching to sample, animals are given a cue or sample (e.g., light) immediately before the trial and must choose the lever associated with the cue. By increasing the interval or delay between the presentation of the cue and the start of the choice trial, memory for the cue can be evaluated. Delayed non-matching to sample employs the same approach except that the animal must choose the lever that does not match the previously displayed cue or sample. Usually, the number of errors, trials to a fixed criterion, the type of errors, and the delay at which criterion responding is no longer maintained are measured.

137. *Differential reinforcement of low rate tasks*: animals must learn to withhold responses for a specified interval before responding to obtain a reinforcer. Responding before the end of the interval restarts the trial. By increasing the interval that the animal must withhold a response, this procedure can be used to measure response inhibition or impulsivity. Usually the delay at which criterion responding is no longer maintained is measured.

138. *Discrimination reversal tasks*: animals must learn that the previously correct response (e.g., right lever or illuminated lever) is no longer the correct lever and that previously incorrect lever is now the correct lever. By measuring the trial to learn the new correct response, this test provides a measure of both learning and response flexibility. Usually, the number of trials to learn the reversal to a set criterion is measured

Postmortem evaluation of nervous system tissues in the Developmental Neurotoxicity (DNT) study

139. The enhanced potential for toxicologic insult to the developing nervous system as compared to the adult, is related to the formative processes (e.g., proliferation, differentiation, migration, restructuring, myelination) that occur in the young. This concept is constant across all human and animal species, although the time scale for critical events in neurodevelopment is species-specific (Moore, 1989; Rice and Barone, 2000). Morphological parameters are evaluated in the DNT study, along with functional and

neurobehavioral endpoints, to assess adverse treatment-related outcomes to the nervous system following developmental exposures.

140. The developmental neurotoxicity testing guideline (TG 426) requires histopathological evaluation of nervous system tissues from selected offspring (10/sex/group) at the time of weaning (around postnatal day 21) or earlier (e.g., PND 11), and at study termination (approximately postnatal day 60). Consequently, samples from weanlings are collected at a time point during active neurological development and during the late stages of the treatment period, while in the adult (PND 60) offspring, samples are collected after the brain has matured and following a period of approximately 40 days during which treatment has not been administered. Following gross pathological examination of nervous system structures and measurement of brain weight (10/sex/group), samples of brain tissue are collected at both ages; peripheral nervous system tissues are collected only in adult offspring. A qualitative assessment of tissues is performed, generally starting with a comparison of control and high-dose subjects, with subsequent evaluation of low- and mid-dose tissues if a treatment-related effect is observed. Subjective analysis of treatment-related findings (conducted without knowledge of treatment group) is also performed. Additionally, a quantitative assessment is conducted at both time points; the guideline specifies measurements of, at a minimum, major regions in the cerebral cortex, hippocampus, and cerebrum.

141. Brain weights are recorded for 10 offspring per sex in each group. Because there are alternative methods for assigning animals to various tests, the offspring assigned for brain weight measurement may or may not also be designated for further histopathological processing and examination. Thus, practical considerations related to the processing of various aged animals at scheduled termination will dictate the fixation status of the test subject brains; examples can be found in TG 426, Appendix 1. Brain weights can be measured either before or after fixation, although the same procedures should be used throughout the study for all groups, and the study report should clearly indicate the procedure used.

142. Gross morphologic examination of the nervous system tissues can be conducted at any point during the processing of tissues for microscopic evaluation.

143. Proper fixation of nervous system tissues is critical to the success of subsequent histopathological evaluation. *In situ* perfusion fixation of adult animals is performed to minimize artifacts introduced into nervous system tissues (OECD, 2004); additionally, it is generally agreed that there is a need for perfusion (versus immersion) fixation of brain tissues in weanling-aged offspring, although there has been some historical controversy over this issue (OECD, 2000). If samples are collected from pups at an earlier age (e.g., at PND 11), immersion fixation is usually considered adequate. However, with any immersion processing, the removal of the calvarium is necessary to ensure adequate contact with the fixative; care should be taken to avoid inadvertent damage of tissues during this procedure. Further discussion of immersion and perfusion fixation techniques can be found in OECD Guidance 20, paragraphs 51-53 (OECD, 2004), as well as in a number of standard published histological protocols (Prophet *et al*., 1994; Fix and Garman, 2000; Bancroft and Gamble, 2002). The fixative used and the length of time that tissues are retained in fixative may be dictated by anticipated histologic procedures such as the use of special staining or immunocytochemical staining (see Garman *et al*., 2001 for a list of special stains and ancillary procedures that can be considered for use in developmental neurotoxicity studies). Prolonged formalin fixation may result in brain shrinkage; therefore, brain slices should be prepared, processed, and embedded at approximately the same time for all animals on which morphometric evaluations may eventually be performed (Garman *et al*., 2001). Paraffin embedding is acceptable for tissues of the central nervous system. For tissues of the peripheral nervous system, the use of osmium post-fixation, together with epoxy embedding, may be useful and appropriate when a higher degree of resolution is required (e.g., for

evaluation of peripheral nerves when a peripheral neuropathy is suspected and/or for morphometric analysis of peripheral nerves).

144. Central and peripheral nervous system structures that should be sampled and examined histopathologically are listed in TG 426. They include all major brain areas (e.g., olfactory bulbs, cerebral cortex, hippocampus, basal ganglia, thalamus, hypothalamus, midbrain [tectum, tegmentum, and cerebral peduncles], pons, medulla oblongata, and cerebellum. Additionally, in adult offspring, representative sections of the spinal cord and the peripheral nervous system should be sampled, including the eye with optic nerve and retina, the spinal cord at the cervical and lumbar swellings, the dorsal and ventral root fibers, the proximal sciatic nerve, the proximal tibial nerve (at the knee), and the tibial nerve calf muscle branches. The spinal cord and peripheral nerve sections should include both cross or transverse and longitudinal sections.

145. It is important that the qualitative neuropathological evaluation should focus on evidence of alterations to normal nervous system development in addition to signs of cellular alteration (e.g., neuronal vacuolation, degeneration, and necrosis) and tissue changes (e.g., gliosis, leukocyte infiltration, and cystic formation) that are typically observed in neuropathology studies in adult animals. Examples of significant alterations that are indicative of developmental insult are detailed in TG 426 and include the following:

- alterations in the gross size or shape of the olfactory bulbs, cerebrum or cerebellum;

- alterations in the relative size of various brain regions, including decreases or increases in the size of regions resulting from the loss or persistence of normally transient populations of cells or axonal projections (e.g., external germinal layer of cerebellum, corpus callosum);

- alterations in proliferation, migration, and differentiation, as indicated by areas of excessive apoptosis or necrosis, clusters or dispersed populations of ectopic, disoriented or malformed neurons or alterations in the relative size of various layers of cortical structures;

- alterations in patterns of myelination, including an overall size reduction or altered staining of myelinated structures;

- evidence of hydrocephalus, in particular enlargement of the ventricles, stenosis of the cerebral aqueduct and thinning of the cerebral hemispheres.

The first procedure in the qualitative evaluation is the comparison of control and high-dose assessments, with subsequent analysis of lower dose groups when necessary, in an attempt to identify a dose-response relationship for any apparent treatment-related findings. Treatment-related findings should also be confirmed through the use of a subjective analysis, where evaluation is conducted without knowledge of treatment group.

146. Quantitative histopathological evaluations (i.e., morphometry data) have been shown to be sensitive endpoints of neurotoxicological effect which cannot be predicted by alterations in qualitative histopathology, brain weight, or body weight (Raffaele *et al*., 2005). Although TG 426 specifies that measurements be taken from major regions in the cerebral cortex, hippocampus, and cerebrum, additional measurements, or measurements that target a specific region of the brain, may be warranted. Careful dissection and tissue preparation are critical to obtaining homologous sections suitable for use in morphometry. Homology of sections (attained through standardized embedding techniques and the use of consistent anatomical landmarks and plane of sectioning) is necessary to obtain meaningful measurements. Various techniques can be used to record measurements, which are typically linear; it is noted that advances in the use of digital technology have increased the level of precision and standardization in quantitative techniques and have allowed easier archiving and verification of data. Alternative methods for

quantitation (e.g., areal measures, cell counts) may provide an approach for further characterizing observed treatment-related morphometric findings (Garman *et al*., 2001).

Data interpretation

Issues in interpretation of neurobehavioral data

147. It is important to recognize the relationships between behavioural measures in the interpretation of behavioural endpoints. For instance, disturbances in motor or sensory function can affect measures of learning and memory if they influence the ability of animals to perform the learning or memory tests.

148. A number of different tests have been used for assessing effects on learning and memory (Table 6). Many of them are based on the animals moving, using their senses, etc., and therefore an impaired performance in a learning test may reflect other behavioural effects than learning impairment. If it is believed that apparently treatment-related impairments on learning or memory evaluations may be due to confounding factors, control procedures or follow-up studies may be used to further elucidate the findings. For example, measurement of swim speed in a straight alley is very useful in distinguishing motor impairment from learning or memory impairments in water maze tasks. It is also important to keep in mind that a measure of memory cannot be interpreted in the absence of a measure of learning obtained in the same test (US EPA 1990).

149. Timing of behavioural assessment should consider the potential of the test substances to produce acute affects on behavioural measures, especially where the test substance is administered directly to offspring daily. Regardless of the route of administration, the ability to distinguish between systemic or developmental effects may be limited when alterations in neurobehavioral parameters are observed during the period of treatment, depending on the kinetics of the compound. In the standard developmental neurotoxicity study design, treatment is stopped following PND 10 or PND 21, while neurobehavioral testing is conducted around the time of weaning, during the time of puberty and young adult growth (functional observational battery only), and again just before termination at approximately PND 60.

150. The apparent reversibility in adult offspring, of effects observed early in life, may be related to compensatory developmental or behavioural processes, and not represent a true recovery. Likewise, findings observed in adult offspring that had not been previously observed in young test subjects should not be discounted for lack of concordance, since they may represent the latent expression of early alterations in neurological development. The interpretation of apparently reversible effects should also be considered. Irreversible effects are clearly serious, while reversible effects may be of concern if exposure is continuous or when acute effects are of concern. It should also be noted that, it is often not possible to determine whether an effect is truly reversible. The nervous system possesses plasticity, which may enable the organism to compensate for compound-related effects, but the resulting alteration in nervous system function should be regarded as an adverse effect. If developmental neurotoxicity is observed only during some portion of the life-span, compensation should be suspected. Also, learning effects observed for example during the beginning of a learning task but not at the end should not be interpreted as reversible effects. Rather the results indicate that the speed of learning is decreased.

151. Consideration should be given to the possible effect of the treatment paradigm on behavioural testing. The timing of testing relative to the timing of daily test substance administration can have profound

effects on the outcome, often dependent upon the pharmacokinetic and pharmacodynamic profile of the chemical. Additionally, it may not be possible to determine whether alterations in neurobehavioral parameters that are observed during the period of treatment are systemic or developmental in nature. Generally, such distinctions can only be made with confidence following the conduct of additional experiments, *e.g.,* pharmacodynamic, mode of action, phased dosing, or cross-fostering studies. This information is typically not considered crucial for the determination of potential risk to susceptible developing populations; however, it may be important within the context of hazard-based chemical labelling.

Test validity for behavioural endpoints

152. Validity refers to what a test measures (relevance) and how well it is measured (reliability). Validation of behavioural test batteries using a potential positive control, *i.e.,* a behavioural teratogen has been done several times. Methyl-mercury and ethanol were investigated in the European Inter-laboratory Study group on Behavioural Teratology with methyl-mercury giving the clearest results (Elsner *et al.,* 1986, Elsner *et al.,* 1988). In a study where propoxyphene, chlorpromazine, and vitamin A were investigated because they produce different behavioural teratogenic profiles, vitamin A was judged as the best positive control in the test battery used (Saillenfait and Vannier 1988).

153. An important aspect of test validity refers to the relevance of the test, *i.e.,* whether a test measures what it is designed to measure. For example, does a change in a learning test always only indicate effects on learning abilities? The answer is no, since amongst others motivation and activity level may influence the performance in a learning test. Therefore, the influence of these factors on the results must be assessed and evaluated before conclusions concerning learning abilities as such can be drawn.

154. The behavioural tests included in test batteries should be individually validated tests and there are several examples of positive control substances in the literature. For example, 2,5-hexanedione can be used as a positive control substance causing effects in Rotarod (Ladefoged *et al.,* 1989). Effects of triethyltin, acrylamide, 2,5-hexanedione, and the lack of effect of insulin, carbon tetrachloride, and haloperidol have been demonstrated in the grip strength apparatus (Gerber and O'Shaughnessy 1986). This shows that the method is sensitive to known neurotoxicants, and will distinguish these from effects on plasma glucose levels, liver function, or CNS dopamine blockade (Crofton *et al*., 2007).

Test Sensitivity

155. The sensitivity of tests is influenced by the biological variation across animals and this is important in behavioural studies. However, many behavioural endpoints are measured more than once for the same animal, *i.e.,* a repeated measures design (Tamura and Buelke-Sam 1992), and when analysing such data with repeated measures analysis of variance, the variation across animals and time is taken into consideration and the sensitivity of the test is increased.

156. The sensitivity of a test relates to the ability to detect changes in the function under evaluation. When evaluating results of behavioural tests it is important to give consideration to the relationship between the complexity of the test and the capabilities of the laboratory animal. An underestimation of the capabilities of the animals means that the test was too easy and had a low sensitivity, *i.e.,* only animals with major defects would not be able to adequately perform the tests. Overestimation means that the control group is unable to correctly complete the task and therefore exposed animals could hardly be worse. For example, it was found in a study of pups pre-natally exposed to xylene that the biological sensitivity of Morris maze was adequate for standard-housed offspring but it was too low (the test was too easy) for enriched-housed animals (Hass *et*

al., 1995). In rats pre-natally exposed to tributyltin and investigated in two spatial learning tasks, *i.e.,* radial arm maze and Morris maze, a clearly retarded acquisition was seen in the radial arm maze while no differences were obtained in the swim maze (Gårdlund *et al.,* 1991). The schedule for testing in Morris maze was, however, very reduced using the same start position for every trial and only one platform position and therefore the sensitivity may have been low.

157. Investigations indicate that many behavioural tests, such as Morris maze, Figure-8 maze, Open field and Rotarod, are able to detect differences around 30-50% from the control value (Hass 1993; Reiter and MacPhail 1982). With a group size of 16 animals, the coefficient of detection, i.e. the percent change required to detect a significant change using a given alpha (*e.g.,* 0.05)**,** was 7-23% for auditory startle habituation, 2-26% for the operant conditioning technique, and 13-20% for a complex water maze procedure (Vorhees 1985). These results indicate that under standardised conditions behavioural measures are as precise as other scientific measures performed on living animals (Annau and Cuomo, 1988).

158. A single positive control cannot, however, be expected to cover all types of behaviour (Saillenfait and Vannier 1988). Therefore, positive control data that demonstrate the sensitivity of each of the tests in a battery can be used and the positive control data do not have to be from studies using prenatal exposures (US EPA 1990, Crofton *et al.*, 2007).

159. The severity and nature of the behavioural effect should be considered. Generally, a pattern of effects (*e.g.,* impaired learning during several consecutive trials) is more persuasive evidence of developmental neurotoxicity than one or a few unrelated changes. However, it is important to note that the different behavioural evaluations included in the guideline are designed to assess different behavioural functions of the animal, thus the failure to find multiple deficits across various types of assessments should not be interpreted as an indication that deficits detected on individual tasks are not related to treatment.

Neuropathological endpoints

160. Neuropathology outcomes are typically presented as incidence data (e.g., for histopathological findings) or as continuous data (e.g., for brain weight and morphometric evaluations). The statistical methods utilized in the data analysis can include analysis of variance (ANOVA), followed by a t-test (e.g., Dunnett's) when data are homogeneous, or the use of non-parametric ANOVA when data are not homogeneous. Alternative methodologies, such as a multivariate analysis of variance may be useful to determine interactions between treatment, gender, and dose. The use of non-parametric methods such as Chi-Square or Fisher's Exact Test can be used to analyze incidence data. Consideration of the litter as the basic statistical unit for evaluation is of paramount importance (Holson and Pearce, 1992), and should be incorporated into both the study design (i.e., during the selection and assignment of animals for testing) and the statistical evaluation of the data. In the statistical assessment of outcome data, it is important to realize that the sample size for brain weight and histopathology evaluation (i.e., 10/sex/group; generally one male or one female from each litter, sometimes representing different cohorts of littermates) provides limited statistical power (Tilson, 1994). If gender-related differences are not observed, power can be increased by combining sexes in the data analyses, since the sample size will then increase from 10 litters per group up to approximately 20. Consultation with a statistician can be useful to confirm appropriate statistical treatment of the data, or to help in resolving difficult or unique analytical issues that arise.

161. As with functional and behavioral endpoints assessed in the DNT study, it is generally not possible to distinguish the precise origin or timing of the toxicological insult that resulted in an observed adverse neuropathological outcome. While comparative studies in adult animals may provide some information in this regard, such data are not always available. For example, no postmortem data are

collected from dams in the DNT study; although there may be some data from adult rats following either acute or repeated exposures in other neurotoxicity studies, exposure during critical windows of development makes the DNT study subjects unique. There may be other factors that render the postmortem data from DNT studies difficult to interpret. For example, in identifying adverse histopathological outcomes, it is important to consider that some developmental events such as programmed cell death may be difficult to distinguish from other treatment-related events with similar morphological characteristics (Shield and Mirkes, 1998). Additionally, disruption of early events in neurological development can alter growth and differentiation, thus producing alterations in neurological structures that may not be associated with cell death or other associated tissue response (Garman *et al*., 2001). Statistically significant differences in group responses may not be detected due to the limited sample size in neuropathology data, and it is important to realize that low incidences of qualitative observations can nevertheless be biologically and/or toxicologically meaningful in spite of a lack of statistical significance. Historical and positive control data from the performing laboratory can be useful in interpreting biological significance. An additional complication in interpreting study results is that findings noted in the neuropathological evaluations conducted on offspring killed at the early time point may or may not be related to later observations. Later observations may represent latent effects of earlier exposure (i.e., during development) and should not be discounted, since they may be potentially related to treatment. However, the absence of observed adverse effects at later time points does not discount earlier findings, nor does it necessarily indicate recovery from adverse developmental effects, since (i) alterations in developmental processes or events observed at an earlier time point may no longer be occurring in adults, (ii) extensive brain remodeling during development, as well as growth of the animal, can mask the observation of some structural anomalies in adult subjects, and (iii) the fact that a different set of animals (albeit originating from the same population of litters) is being examined at each time point can provide a source of potential variability in response. It is also important to note that treatment-related observations in the adult offspring are not likely related to degenerative processes that may result from normal aging (such as might be observed after 2 years of treatment in a chronic rodent study), since the DNT study animals are young adults (2 months of age) at the time of termination. And finally, neuropathological alterations may be independent of treatment-related behavioral outcomes (Jensen and Catalano, 1998). The evaluation should include the relationship, if any, between observed neuropathological and behavioural alterations (OECD, 2007a).

Relationship of Maternal and Offspring End Points

162. There may be a relationship between maternal and offspring end points, especially during the lactation period where the offspring are depending on the maternal care. The relationship can be documented by the use of cross-fostering of treated offspring to untreated mothers and untreated offspring to treated mothers. In any case, whether the effect is mediated via the maternal toxicity or not, it should be regarded when setting the NOAEL for developmental toxicity.

Litter Size, Sex and Mean Pup Body Weight

163. The effect of litter size on pup body weight is well established. In general, pup body weight is inversely related to litter size at least in pre-weaning animals. Though this relationship may disappear after weaning statistical analysis of data such as pup body weight should adjust for litter size where litter size in not normally standardized (i.e., cull litters to standard size). Gender also influences body weight and size. Measures that may be indirectly related to weight or size should consider the potential role of body weight or size in explaining gender differences. In addition, where relevant tests should be modified to adjust for differences in size between genders (e.g., beam width in beam test). Males are generally slightly heavier than females in the pre-weaning period and the difference increases markedly around and after sexual maturation. If both male and female offspring are included in the statistical analysis, sex should be included in the analysis. In general, statistically significant changes in offspring body weights are considered as adverse.

Physical and Functional Developmental Landmarks

164. Delays of physical developmental and reflex ontogeny in the lactation period often occur in the presence of decreased body weight of the offspring and may as such be signs of developmental delay. In cases where changes in physical or functional development are observed in exposed pups without a corresponding change in growth, the changes may reflect specific developmental effects, and not a general delay in development.

165. AGD may be influenced by the size of the animal and this should be taken into account when evaluating the data. The size or length of the pups is normally not measured (sometimes crown-rump), but body weights are measured. In some cases, the anogenital index, *i.e.,* AGD divided by body weight, is used. However, body weights of pups may be quite variable leading to a large variation in the anogenital index. This could mask eventual effects on AGD and is therefore not recommended. Instead, the size of the animals should be accounted for by including a covariant. Body weight can be used, but this parameter is in three dimensions, while AGD is in one dimension. Consequently, the optimal co-variate seems to be the cube root of the body weight (Clark, 1999). A statistically significant change in AGD that cannot be explained by the size of the animal indicates effects of the exposure and should be used for setting the NOAEL.

IV. REPRODUCTIVE ENDPOINTS IN ADULTS

Overview of Endpoints Assessed

166. A number of Test Guidelines is available that includes an assessment of reproductive endpoints including the "Reproduction/Developmental Toxicity Screening Study" (TG 421), the combined "Repeated Dose Toxicity Study and Reproduction/Developmental Toxicity Screening Study" (TG 422), the "One-Generation Reproductive Toxicity Study" (TG 415) and the "Two-Generation Reproductive Toxicity Study" (TG 416). All of these studies provide some information on the adult reproductive system following exposures to mature animals. In addition, the two-generation reproductive toxicity study provides information on the reproductive system following exposures during all phases of development; thus, the F1 generation is exposed *in utero,* during lactation, and continuing through sexual maturation into adulthood. A recent OECD draft guidance document includes information on histopathologic evaluation of male and female reproductive tissues and vaginal smears and may be be very useful when assessing and interpreting these studies (OECD, 2008).

Methodological Issues

Examination of Male Reproductive Organs

167. A macroscopic evaluation of the male reproductive organs is included in most reproductive toxicity Test Guidelines. Absolute and relative (*i.e.,* adjusted for body weight) weights of the testes, epididymides, seminal vesicles, prostate and pituitary are recorded. A number of reviews are available outlining the proper techniques for the histological examination of the testes (Russell *et al.,* 1990; U.S. EPA, 1996; Chapin and Conner, 1999; Lanning *et al*, 2002; Creasy, 2003; Latendresse *et al*., 2002). The method used to examine the histology of the testis and epididymis is dependent on the specific reproductive toxicity study protocol that was followed. In protocols that specify dosing of the male for an entire course of spermatogenesis combined with analysis of the sperm (*e.g*., TG 416), less extensive histological examination is required than in situations where the dosing regimen is shorter and no sperm analyses are conducted (*e.g.,* TG 421, 422). For the one-generation (TG 415) and two-generation reproductive toxicity studies (TG 416), the testis and epididymis, including the caput, corpus and cauda segments, are examined histologically following appropriate fixation and sectioning. Several aspects of testicular degeneration can be recognised including hypocellularity (decreased numbers of germ cells in the epithelium), vacuolation of the Sertoli cells, formation of multinucleated cells (giant cells, composed of spermatocytes and spermatids), cell death and unusual associations of germ cell stages within the seminiferous tubules, and the presence of round spermatids and cellular debris in the epididymis.

168. For shorter-term studies (*e.g.,* TG 421, 422), it may be possible to examine potential injury to the testis at specific stages of the cycle of the seminiferous epithelium. It is important to also consider the age of the animals at the initiation of dosing to ensure that sexual maturity had been attained. The most commonly used system for classifying the 14 stages in the rat was described by Leblond and Clermont (1952) based on the development of the acrosome and the shape of the elongate spermatid nucleus. For standard toxicology studies, it is adequate to simply assess qualitatively whether all stages of spermatogenesis are normal, and a quantitative categorization of the stage of each tubular cross section of the testis is not necessary (OECD, 2008).

Sperm Parameters

169. The parameters that are included in the two-generation reproductive toxicity study (TG 416) are sperm number, sperm morphology, and sperm motility. A number of reviews are available outlining the proper techniques for sperm analyses (US EPA, 1996; Seed *et al.,* 1996; Chapin and Conner, 1999). For analysis of sperm motility, samples can be collected from the cauda epididymidis or the vas deferens. The samples can be collected anywhere from room temperature to 37ºC, and can be stored in any physiologic buffered saline solution for up to one hour. For sampling from the cauda, the tissue is placed in a dish with an aliquot of buffer and the cauda is nicked in a few sites with a blade. The sperm then diffuse into the medium and the tissue is removed. For sampling from the vas deferens, a small amount of tissue is placed in a dish with buffer and the sperm diffuse into the medium. In sampling from either tissue, care should be taken to avoid any unnecessary manipulation of the tissue.

170. Sperm motility can be assessed manually or by computer-assisted sperm analysis (CASA) systems. Sperm motility should be measured prior to morphology or count because motility is critically dependent on temperature. When assessing sperm motility, the samples should be at a temperature of 34-37ºC. The depth of the chamber is critical for an accurate assessment of motility; chambers greater than 20 μm are preferable for rodents. In general, 200 sperm should be analysed. A minimum value of 70% motility is acceptable in controls. For manual assessments, the number of motile sperm is counted with a hema-cytometer. One easy method is to count the number of stationary sperm, fix the sample and then count the number of total sperm. For CASA analysis, a sperm is considered motile if the average path velocity is greater than a user-defined threshold, the threshold is determined for a given laboratory.

171. In addition, the percentage of progressively motile sperm is assessed. This measure distinguishes sperm that are simply twitching in place from those that are making forward progress. This can be done during manual assessments. For CASA analysis, progressive motility is defined as sperm that exceed user-defined thresholds for one or more CASA outcomes, usually average path velocity and straightness, or straightline velocity alone (Seed *et al*., 1996; Perreault and Cancel, 2001).

172. Sperm morphology can be assessed from samples from the cauda epididymidis or the vas deferens. Generally, sperm morphology is assessed from a sample that has not been collected for the assessment of sperm motility as the bovine serum albumin that is present in the buffer can interfere with the stains that are used to assess morphology. For the assessment of morphology, a small sample is placed on a slide and can be viewed either as a wet preparation or the slide can be air-dried. Drying can cause kinking of the tails, therefore some laboratories prefer viewing wet samples. Typically the samples are stained with Eosin Y, but a variety of stains are acceptable as long as they allow the appropriate viewing of the sperm. The samples can be viewed with a light microscope at a magnification of 400X.

173. There is no universal classification scheme for sperm morphology. Sperm abnormalities are generally described in terms of abnormalities of the head and tail, and can include a head that has too little or too much hook and a tail that is frayed or coiled, multiple tails, misplaced mitochondria, or there can be a residual drop of cytoplasm remaining on the tail. In addition, the neck at the junction of the head and tail can be a weak spot. It is often helpful to present the incidences both including and excluding tail abnormalities, since the relatively larger number of tail abnormalities may mask an increase in other abnormalities. In rodents, there is a high proportion of normal sperm so usually 200-400 sperm per rat are evaluated. The statistical power to detect differences in the percentage of normal sperm between groups of control and treated rats is high.

174. In the two-generation reproductive toxicity study (TG 416), cauda epididymal and testicular sperm counts are recorded. The cauda sperm count is a measure of sperm storage. Testicular sperm counts generally refer to enumeration of detergent and homogenisation-resistant spermatid heads. During spermiogenesis, the nuclei of the spermatids become highly condensed. The mature spermatid nuclei, unlike less mature spermatids, are very resistant to homogenisation. Thus, this technique provides a reliable estimate of the number of spermatids in the later stages of spermiogenesis. Sperm counts can be determined at the time of necropsy or the tissue can be frozen for later analysis. For both cauda epididymal and testicular counts, the tissue is homogenised in the presence of detergent such as Triton X-100®. Sperm counts can be obtained by use of a hema-cytometer, coulter counter, or CASA. Each laboratory should determine the variability associated with the counts and the level of statistical power.

Examination of Female Reproductive Organs

175. The female reproductive organs are examined macroscopically in all reproductive Test Guidelines (TG's 414, 421, 422, 415 and 416). The number of implantation sites in the uterus is recorded. The absolute and relative (*i.e.,* adjusted for body weight) weights of the ovary, uterus, vagina and pituitary are recorded, and these organs are examined for histopathology. Detailed histological examination of the ovaries should cover the follicular, luteal, and interstitial compartments of the ovary, as well as the epithelial capsule and ovarian stroma.

Oocyte Quantitation

176. In the two generation reproductive toxicity study, a quantitative evaluation of the primordial follicles is conducted for the F1 females. The methods for appropriate fixation and staining, and quantitative analysis have recently been reviewed by Heindel (1999). The follicles can be categorised into three classes.

- Nongrowing, primodial follicles are defined as an isolated oocyte or an oocyte that is surrounded by a partial or unbroken single layer of granulosa cells.
- Growing follicles are defined as an oocyte that is surrounded by a multi-layered, solid mantle of granulosa cells without evidence of a fluid-filled cavity (antrum), and,
- antral follicles are characterized by a central oocyte and a fluid-filled antrum bordered by hundreds of layered granulosa cells.

177. For the TG 416, a standard procedure is to evaluate 10 females in a treatment group, and to count the follicles in every tenth sequential section beginning with the section when follicles are first noted. However, some researchers have shown that the same statistical power can be achieved by counting fewer sections and/or by evaluating fewer animals. Each laboratory should determine the appropriate number of animals and number of sections to evaluate.

Vaginal Cytology

178. Vaginal cytology is evaluated to determine the length and normality of the oestrus cycle in the P and F1 females in the two-generation reproductive toxicity study. A recent review of the methods has been conducted by Cooper and Goldman (1999). OECD (2008) also provides guidance on collection and interpretation of vaginal smears. Vaginal smears must be collected daily for at least two weeks for an

accurate determination of cycle length. There are several methods available for collection of the smears. One method involves collection by lavaging the vagina with an eyedropper containing approximately 0.25 mL water or physiological saline and drawing the fluid back into the eyedropper. Only the tip of the eyedropper is inserted into the vagina; stimulation of the cervix should be avoided to ensure that pseudo-pregnancy is not induced by the method. The fluid is expelled evenly onto a microscope slide. The smears can be viewed immediately at approximately 100X magnification, and either stained for subsequent analysis or viewed wet and stored for subsequent staining. Vaginal smears are typically stained (*e.g.*, 1% aqueous toluidine blue O, methylene blue).

179. The entire smear should be examined as the cell distribution may be uneven. The smears are evaluated based on the types of cells present. In the rat, the oestrus cycle lasts for 4-5 days. Each cycle has three distinct phases including met-oestrus/di-oestrus (2-3 days), pro-oestrus (1 day), and oestrus (1 day). During di-oestrus, the smear contains a mixture of cell types; leukocytes are the predominant cell type with a varied number of cornified epithelial cells. During pro-oestrus, the smear contains a predominance of clumps of round, nucleated epithelial cells, and during oestrus, the smear contains mainly cornified cells. Different laboratories may classify the smears according to different criteria, and as such it is important that the females are monitored over an extended time for comparison of the smears with the cycling status of the females. In addition, it is important that the smears are collected at the same time each day to reduce variability and enable identification of cycling patterns.

Reproductive Performance

180. Reproductive performance is the ability of male and female animals to successfully mate and produce viable offspring, and is assessed in the TG 421, 422, 415 and 416. Information on reproductive performance is generally expressed as indices that are ratios derived from the data collected in the studies. Many indices are used and may be laboratory specific, but the major indices are described in the following table.

Table 7 Calculation of reproductive indices

Index	Calculation	Definition
Male Mating Index	$\frac{\text{No. of males with confirmed mating}}{\text{Total No. of males cohabited}} \times 100$	Measure of male's ability to mate
Female Mating Index	$\frac{\text{No. of sperm-positive females}}{\text{Total No. of females cohabited}} \times 100$	Measure of female's ability to mate
Male Fertility Index	$\frac{\text{No. of males impregnating a female}}{\text{Total No. of males cohabited}} \times 100$	Measure of male's ability to produce sperm that can fertilise eggs
Female Fertility Index	$\frac{\text{No. of pregnant females}}{\text{No. of sperm-positive females}} \times 100$	Measure of female's ability to become pregnant
Gestation Index	$\frac{\text{No. of females with live born pups}}{\text{No. of pregnant females}} \times 100$	Measure of pregnancy that provides at least one live pup
Survival Index	$\frac{\text{No. of live pups (at designated time)}}{\text{No. of pups born}} \times 100$	Measure of pup survival which is calculated at several times during lactation

Data Interpretation

Male Reproductive Organs

181. Both absolute and relative weights of the male reproductive organs should be considered as a decrease in absolute weight may occur and may not necessarily be related to a reduction in body weight gain. However, care should be taken in interpreting data where a substantial bodyweight effect is evident. Since there is low inter-animal variability in testis weight, a significant change in absolute testis weight (increase or decrease) can indicate an adverse effect. Changes in testis weight can be due to damage to the seminiferous tubules or a variety of other cases including oedema, inflammation, cellular infiltration, Leydig cell hyper-plasia or fluid accumulation due to blocked efferent ducts. The weights of the prostate and seminal vesicles are androgen dependent and therefore may reflect changes in testicular function or endocrine status. Pituitary gland weight can also be an indicator of reproductive status. However, the pituitary contains many cell types that are responsible for the regulation of several physiologic functions including some that are separate from reproduction. Therefore, gonadotrophin-specific histopathological evaluation may be useful to determine which cell types are affected.

182. Histopathological evaluations of the reproductive tissues are relatively sensitive indicators of damage and are valuable for the assessment of male reproductive toxicity. Histopathological findings are generally classified according to qualitative criteria and the data are presented as the number of animals affected within a dose group. There is no standardised method of further quantifying the extent of the damage. In general, histopathological findings should be considered in light of the specific test protocol and in conjunction with information on sperm parameters and reproductive performance. Depending on the test protocol that was used, and the extent of the damage, there may not be an obvious relationship between the histopathological findings and fertility. For short-term studies (*i.e.,* TG 421 and 422) in which the animals are treated for less than the duration of the spermatogenic cycle, an effect on spermatogenesis may not have had adequate time to become evident as reduced sperm counts that affect fertility. In general, any dose related significant histopathological change indicates that there is a potential for similar effect in humans.

Sperm Parameters

183. Two measures of sperm motility are usually calculated. The percentage of motile sperm is defined as the number of motile sperm/total number of sperm X100, and the percentage of progressively motile sperm is defined as the number of progressively motile sperm/total number of sperm X100. Studies have shown that there is a relationship between sperm motility and fertility, but there is no generally accepted standard of how much of a change in motility should be considered adverse. The specific protocol should be considered, as well as information from the other sperm parameters and histopathology. Since sperm motility is dependent on testicular and epididymal function, damage to the testis, as revealed by the histologic appearance, may lead to changes in sperm motility. However, motility can be altered by direct effects on the epididymides or the sperm themselves. Therefore, the absence of a histological lesion does not necessarily mean that a change in motility should be discounted. A dose-response trend and a statistically significant change in sperm motility would generally be interpreted as indicating a potential effect on fertility in humans.

184. Sperm morphology is related to fertility, but there is no generally accepted standard of how much of a change in morphology should be considered adverse. Information on sperm motility and count, as well as histopathology should be considered in the interpretation of sperm morphology. Histological lesions of sufficient magnitude can impact sperm morphology. However, normal sperm morphology is dependent on numerous factors including the correct assembly of protamines, so changes in sperm morphology should not be discounted in the absence of histological lesions.

185. Sperm count data are usually expressed in two ways. Information regarding the number of sperm available for ejaculation is provided by the number of sperm per cauda and the number of homogenisation-resistant spermatids per testis. Information regarding the efficiency of the tissue is provided by the number of sperm per milligram cauda and the number of homogenisation-resistant spermatids per milligram testis. Studies have demonstrated a strong relationship between sperm counts and fertility in all species that have been examined. Again, information on the other sperm parameters and histopatholgy should be considered in the overall interpretation. Testicular lesions of sufficient magnitude will be reflected in the sperm counts, but changes in sperm counts should not be discounted in the absence of histological lesions. Similarly, a reduction in sperm count may not result in reduced fertility, particularly in rodent studies. This is due to the fact that rats and mice have a tremendous excess of spermatozoa in their ejaculates, and as such sperm counts have to be reduced by as much as 90% to affect fertility. It is important to note that sperm concentrations in human males are highly variable and generally lower than in rodents. The distribution of counts is such that many men have sperm concentrations near or below WHO reference values for fertility. Therefore, even a small decrease in sperm concentration across a population would be expected to shift the fertility potential of the group and move some men into the infertile or subfertile range. For this reason, a statistically significant change in sperm count in a rodent study is considered to be indicative of a potential effect on fertility in humans.

Female Reproductive Organs

186. The information on the weights and histopathology of the female reproductive organs as well as reproductive performance should be evaluated together, with consideration of the specific test protocol. Uterine weight fluctuates 3- to 4-fold during the oestrus cycle, peaking at pro-oestrus when it is filled with fluid in response to increased oestrogen secretion. Thus, compounds that inhibit steroidogenesis and cyclicity can cause the uterus to become atrophic; conversely compounds that are estrogenic can cause the uterus to become quite large. Uterine weights taken from cycling animals have a high variance, and only large compound-related effects will be demonstrated unless the data are interpreted relative to the animal's oestrus cycle stage. Similarly, the histologic appearance of the uterus varies with the stage of the oestrus cycle and pregnancy. The uterine endometrium, which is sensitive to estrogens and progestogens, will show hypertrophy and hyperplasia in response to these kinds of compounds, and will show hypoplasia and atrophy in response to compounds that inhibit steroidogenesis. Effects induced during development can result in delayed puberty and persistence of infantile genitalia. Information on the number of implantation sites should be interpreted along with the information on live pups and resorptions as described in paragraph 64.

187. A limited amount of information regarding ovarian histopathology can be obtained from female rodents that have been treated according to Test Guidelines 421 or 422 as the dams are sacrificed on lactation day 4, and therefore are not actively cycling. In the rat, ovarian weight does not fluctuate during the oestrus cycle, and any changes should be considered adverse. The function of the ovary shifts during the oestrus cycle so histopathology can reveal a variety of effects including oocyte and follicle depletion, persistent polycystic ovaries, inhibition of corpus luteum formation and luteal cyst development. It is important to note that not all histological alterations will affect ovarian weight, so the lack of an effect on

ovarian weight does not preclude the need for histological evaluation. Similarly, the nature and the magnitude of the histological lesion will determine whether there is a concomitant effect on reproductive performance. In general, any dose related significant histological finding would be considered to indicate that there is a potential for similar effects in humans (OECD, 2008).

188. Many studies have shown a relationship between follicle number and fertility. The relationship between follicle number and onset of menopause is clear in humans but the connection with reproductive senescence in rodents is less clear cut. There is no generally accepted standard of how much of a change in follicle counts should be considered adverse. The information on follicle counts should be examined in conjunction with the histological information and reproductive performance. Often a change in follicle number will be apparent prior to a change in organ weight or histopathology. The magnitude of the reduction in the number of follicles will determine whether there is an effect on reproductive performance. A decrease in follicle counts could indicate either direct oocyte toxicity, or an effect on the granulosa or thecal cells that alters the paracrine control of folliculogenesis. A dose-response trend and a statistically significant change in follicle number would indicate a potential effect in humans.

189. Vaginal weight changes should parallel those seen in the uterus during the oestrus cycle, although the magnitude of the change will be smaller. The vaginal smear data collected in the two-generation reproductive toxicity study can provide information on cycle length, persistence of oestrus, persistence of di-oestrus, incidence of pseudo-pregnancy. Cycle length is determined by selecting a stage in the cycle and counting the number of days until that stage reoccurs. For statistical analysis, comparisons can be made among animals exhibiting normal 4-5 day cycles and abnormal cycles. Alternatively, altered cyclicity may be reflected by the percentage of time spent in one stage of the cycle, and comparisons can be made based on the number of days spent in di-oestrus or oestrus. However, this kind of analysis must take into account individual cycling patterns as it is possible that more than one stage may be affected; analyses that focus on just one stage may mask such effects. An effect on the oestrus cycle can also impact reproductive performance, but this will depend on the nature and magnitude of the oestrus cycle effect. In general, a statistically significant change in the length of the cycle or prolonged oestrus or di-oestrus would be considered adverse. Effects induced during development can lead to agenesis, hypoplasia, and dysgenesis of the vagina. Hypoplasia of the vagina can be concomitant with hyperplasia of the external genitalia, and altered AGD. The opening of the vaginal orifice at puberty is a simple and useful developmental marker.

190. Alterations in pituitary weight in female rodents are indications of an adverse effect. Increased pituitary weight in rodents due to exposure to estrogenic compounds often precedes tumour formation, and may be accompanied by hyper-prolactinemia and a persistent vaginal cornified smear pattern. Decreased pituitary weight due to decreased estrogenic stimulation is less common. The discussion on pituitary weight and histopathology in males is also applicable to females.

Reproductive Performance

191. The data on reproductive performance has to be interpreted with reference to the specific protocol that was used. The data will not necessarily be similar for studies in which there is more than one mating per generation and more than one breeding generation. In addition, in the TG 421, 422, 415 and 416 both sexes are treated, and therefore it is usually not possible to determine whether an effect on some aspect of reproductive performance is due to the female or the male or both. As noted in the preceding sections, it is also important to evaluate the various indices of reproductive performance along with other available information including histopathology, sperm parameters, follicle numbers, and oestrus cycling.

192. The mating index provides information on the integrated function of the neuroendocrine-gonadal axis. It can be affected by many factors including libido, hormonal imbalance and oestrus cycle disruptions. In most protocols, it will not be possible to determine the cause of an effect on mating since mating behaviour can be impacted by lesions in the nervous system or by changes in hormones.

193. The fertility index provides information on the ability of the male and female to achieve a pregnancy. The interpretation of the fertility data should consider the protocol that was used and in particular the duration of treatment of the males prior to mating. In protocols in which the males are treated for less than the duration of the spermatogenic cycle (*i.e.,* TG 421 and 422), a reproductive effect may not be manifested in the fertility index. When there is an effect on the fertility index, it may be difficult to determine the affected sex in studies where both sexes are dosed. The fertility information should be considered in conjunction with the available information on histopathology, sperm parameters, follicle numbers and oestrus cycling. As noted above, the male rodent has a large excess of spermatazoa and therefore it takes a large reduction in sperm number to be reflected as a change in the fertility index. Thus, the fertility index alone can be a rather insensitive endpoint.

194. The gestation index should be treated cautiously. It is not a particularly sensitive endpoint since all litters, regardless of size, are treated equally. Therefore, the gestation index will not provide information on an increased incidence of resorptions.

195. The length of gestation and duration of parturition should be evaluated in conjunction with information on the birth weights and pup survival. A significant reduction in gestation length may also result in reduced birth weights and pup survival. A significant increase in gestation length can also result in a difficult delivery (dystocia) that may impact the health of the dam and/or pups. In addition, the birth weights of the pups may be higher than normal due to extended intrauterine development. In situations where there is a small litter size and/or marked in-utero growth retardation, the gestation length may be extended slightly so that the offspring are born at a similar body weight as their control counterparts.

196. Litter size is an important indicator of overall reproductive performance. A decrease in litter size can result from several factors including decreases in the number of oocytes ovulated, failed fertilization of those oocytes due to poor sperm measures (counts, motility, morphology), and/or increased pre- or post-implantation embryo or fetal loss. Thus, the litter size information should also be interpreted in conjunction with the other available information on reproductive endpoints.

197. Changes in the sex ratio index can result from a number of factors including selective changes in the production of X and Y sperm resulting in different proportions of male or female conceptuses, with no change in litter size, or selective loss of male or female fetuses resulting in small litters. Altered hormone status can also masculinize a female fetus or feminize a male fetus leading to misclassification of gender at birth (without affecting litter size). Thus, the sex ratio information should be evaluated in conjunction with information on embryonic and foetal loss, as well as information from genetic toxicity studies. In addition, a change in the sex ratio can indicate that the chemical has endocrine activity and may have nothing to do with XY alterations.

198. The survival index reflects the ability of the pups to survive and is an important endpoint. The pups will begin to consume food and water on lactation day 14 and therefore may also directly consume the test substance. Reduced survival can result from a number of factors including developmental effects in the pups from either prenatal or lactational exposures, structural abnormalities, maternal neglect, or insufficient milk. Interpretation of the information on survival should also consider the information on pre- and post-implantation loss, number of stillborn pups, total number of dead pups, the number of affected pups per litter, and the pattern of mortality.

V. DATA GAPS

199. Test Guidelines have been designed as tests for hazard identification and dose response assessment, and as such, they provide a valuable tool for risk assessment. The studies described in this document facilitate the broad assessment of chemicals for potential developmental and reproductive toxicity. However, the sensitivity of the studies in the detection of effects is inherently limited by the protocols and how they are utilized. Some of these limitations, described below, may be important to consider when interpreting study results.

Endpoints

200. The approach used in Test Guidelines for reproduction and developmental toxicity testing does not incorporate chemical-specific information, such as mode of action or known target organ specificity, into the design of the studies or the endpoints assessed.

201. An understanding of the pharmacokinetic and pharmacodynamic profile of a test substance in the developing system and of the complexities of direct and indirect developmental exposures during pregnancy, lactation, and to neonates by various routes of exposure is critical to study design, dose selection, and the interpretation and extrapolation of reproductive toxicity data. Current Test Guidelines do not address the collection or use of these data.

202. Current developmental and reproductive toxicity Test Guidelines provide only limited focus on alterations in the ontogeny of organ system function (US EPA, 2002). While reproductive system and nervous system functions are extensively assessed in the multi-generation reproduction study and the developmental neurotoxicity study, respectively, other functional endpoints are not included. For example, the development of normal functional capabilities of the cardiac, urinary, respiratory, immune, and endocrine systems is not assessed.

203. No specific standardised Test Guideline exists for second tier testing on chemicals that are suspected to enhance carcinogenic response following perinatal exposure. The NTP however, has now moved to having perinatal exposure in its chronic/ carcinogenicity studies as its default exposure regime.

204. Neurobehavioral and neuropathological assessments included in the developmental neurotoxicity guideline provide a broad screen for apical effects on multiple functional domains. However, primarily as a function of the animal model typically used, the study does not address some aspects of behaviour (*e.g.*, social behaviour), and cannot provide information on higher level cortical functions that have serious implications in humans. Additionally, specialised follow-up on developmental neurotoxicity testing is not addressed. These would include experiments in which relatively sensitive measures of sensory and/or cognitive function are evaluated in the offspring of animals exposed to chemicals during pregnancy and/or postnatally.

Critical windows of exposure and effect

205. While current guideline studies provide opportunities for exposure across multiple periods of development, the studies are not designed to determine the critical window of exposure. Such information could be useful in targeting a risk assessment to various susceptible population subgroups.

206. With the exception of the developmental neurotoxicity study, which includes a period (approximately 40 days) without treatment prior to neurobehavioral and neuropathological evaluations of offspring at study termination, the issue of longer-term consequences of developmental exposures is not examined. While the multi-generation reproduction study maintains the study animals in each generation to an age of reproductive maturity, the animals are treated throughout the duration of the study, thereby limiting the ability to determine the extent to which adverse treatment-related effects observed during postnatal and adult life stages can be attributed to developmental versus systemic toxicity.

207. Additionally, none of the current Test Guidelines address the potential for effects of developmental exposures in aging individuals. In the prenatal developmental toxicity study, treated offspring are killed immediately prior to birth; in the single- or multi-generation reproduction study, treated adults are maintained on study only through the reproduction phases and killed at approximately 6-9 months of age, while offspring that are not retained for reproduction purposes are killed at weaning; in the developmental neurotoxicity study, treated offspring are maintained on study only until approximately postnatal day 70. Typical Test Guideline rodent chronic/carcinogenicity studies initiate treatment in young adult animals (5-6 weeks of age); therefore, the examinations conducted in aged animals on these studies after 18-24 months of treatment are not relevant to assessing potential effects of exposures during either pre- or postnatal development. The need for expanded assessment of risks to the aged has been widely recognized, and a number of challenges in assessing these risks have been identified (Geller and Zenick, 2005). While standard study designs do not address potential effects of exposures during development on adverse outcomes in aged animals, multigeneration and developmental neurotoxicity study designs could potentially be altered (e.g., by maintaining a cohort of developmentally-treated F1 animals for toxicological assessment at an advance life stage) to address such issues. Although there could be significant challenges in conducting such a study, and in interpreting the resulting data, the knowledge and insights gained from conducting such a study might be substantial and invaluable to the assessment of risk.

VI. REFERENCES

Adams, J. (1986) Methods in behavioral teratology. Handbook of behavioral teratology, Riley EP, Vorhees CV (eds): New York: Plenum Press, 67-97.

Akaike, M., Tanaka, K., Goto, M. and T. Sakaguchi, (1988) Impaired Biel and Radial arm maze learning in rats with methylnitrosurea-induced microcephaly. Neurotoxicol. Teratol. 10: 327-332.

Akaike, M., Ohno, H., Tsutsumi, S. and M. Omusu, (1994) Comparison of four spatial maze learning tests with methylnitrosourea-induced microcephaly rats. Teratology 49: 83-89.

Altman, J. and K. Sudarshan, (1975) Postnatal development of locomotion in the laboratory rat. Anim. Behav. 23: 896-920.

Annau, Z. and V. Cuomo, (1988) Mechanisms of neurotoxicity and their relationship to behavioral changes. Toxicology 49: 219-225.

Anon (1967) European Commission Directive 67/548/EEC (Annex VI).

Anon (2001) European Commission Directive 67/548/EEC, updated by Commission Directive 2001/59/EC.

ATLA (2002) Alternative non-animal methods for chemical testing: current status and future prospects. 10. Reproductive toxicity. (Eds.) Worth, A.P., Balls, M., 30 (Supplement 1), 95-102.

Bammer, G. (1982) Pharmacological investigations of neurotransmitter involvement in passive avoidance responding: A review and some new results. Neurosci. Behav. Rev., 6:247-296.

Bancroft, J.D., Gamble, M. (2002) *Theory and Practice of Histological Techniques*, 5th edition, Churchill Livingstone, London.

Barlow, S. M. and F. M. Sullivan, (1975) 6. Behavioural teratology. Teratology trends and applications, Berry C. L., Poswillo D. E. (eds): Springer Verlag, 103-120.

Barlow, S. M. and F. M. Sullivan (1982) Reproductive toxicity testing in animals in Reproductive Hazards of Industrial Chemicals. An evaluation of animal and human data. Academic Press INC. (London LTD)).

Blanchard, B. A., Riley, E. P. and J. H. Hannigan, (1987) Deficits on a Spatial Navigation Task Following Prenatal Exposure to Ethanol. Neurotoxicol Teratol. 9: 253-258.

Bogo, V., Hill, T. A. and R. W. Young, (1981) Comparison of accelerod and rotarod sensitivity in detecting ethanol- and acrylamide-induced performance decrements in rats: Review of experimental considerations of rotating rod systems. Neurotoxicol. 2: 765-787.

Brandeis, R., Brandys, Y., Yehuda, S. (1989) The use of the Morris water maze in the study of memory and learning. Int. J. Neurosci., 48:29-69.

Brian JV, Harris CA, Scholze M, Backhaus T, Booy P, Lamoree M, Pojana G, Jonkers N, Runnalls T, Bonfà A, Marcomini A, Sumpter JP (2005). Accurate prediction of the response of freshwater fish to a mixture of estrogenic chemicals. Environ Health Perspect. 113(6):721-8.

Buelke-Sam, J. and C. A. Kimmel, (1979) Development and standardization of screening methods for behavioural teratology. Teratology, 20: 17-29.

Buelke-Sam, J., C. A. Kimmel, J. Adams, C. J. Nelson, C. V. Vorhees, D. C. Wright, V. St.Omer, B. A. Koral, R. E. Butcher, M. A. Geyer, J. F. Holson, C. L. Kutcher, and M. J. Wayner. (1985) Collaborative behavioral teratology study: Results. Neurobehav Toxicol Teratol 7:591-624.

Bushnell, P. J., Kelly, K. L. and K. M. Crofton, (1994) Effects of toluene inhalation on detection of auditory signals in rats. Neurotox Teratol. 16: 149-160.

Calabrese E. J. (1983) Animal Extrapolation, chapter 14 in Principles of Animal Extrapolation, John Wiley and Sons.

Cappon GD, Fleeman RE, Chapin RE, and Hurtt ME. (2005) Effects of feed restrictions during organogenesis on embryo-fetal development in rabbit. Birth Defects Research (Part B) 74: 442-430.

Chahoud, I., Buschmann, J., Clark, R., Druga, A., Falke, H., Faqi, A., Hansen, E., Heinrich-Hirsch, B., Hellwig, J., Lingk, W., Parkinson, M., Paumgartten, F.J., Pfeil, R., Platzek, T., Scialli, A.R., Seed, J., Stahlmann, R., Ulbrich, B., Wu, X., Yasda, M., Younes, M, and Solecki, R. (1999) Classification terms in developmental toxicology: need for harmonization. Report of the Second Workshop on the Terminology in Developmental Toxicology, Berlin, 27-28 August 1998. Reprod Toxicol. 13(1): 77-82.

Chapin, R. E., Sloane, R. A., and J. K. Haseman. (1997) The relationships among reproductive endpoints in Swiss mice, using the reproductive assessment by Continous Breeding database. Fundam Appl Toxicol. 38(2):129-42.

Chapin, R.E. and R.A. Sloane. (1997) Reproductive assessment by continuous breeding: evolving study design and summaries of ninety studies. Environ Health Perspect 105(Suppl. 1): 199-205. Chapin, R.E., Sloane, R.A. and J.K. Haseman. (1998) Reproductive endpoints in general toxicity studies: are they predictive? Reprod Toxicol 12: 489-494.

Chapin, R. E. and M. W. Conner, (1999) Testicular histology and sperm parameters. In: An evaluation and interpretation of reproductive endpoints for human health risk assessment. Eds: G. Daston and C. Kimmel, International Life Sciences Institute Press.

Clark, R.L. (1999) Endpoints of reproductive system development. In: An evaluation and interpretation of reproductive endpoints for human health risk assessment. Eds: G. Daston and C. Kimmel, International Life Sciences Institute Press.

Clemens, G. R., Petrere, J. A., and K. Oberholtzer. (1994) Historical control database survey (HCDS) Phase II (P2): External and visceral malformations in the Sprague-Dawley rat and the New Zealand White rabbit. Teratology 49:388-9.

Cooper, R. L. and J. M. Goldman. (1999). Vaginal cytology. In: An evaluation and interpretation of reproductive endpoints for human health risk assessment. Eds: G. Daston and C. Kimmel, International Life Sciences Institute Press.

Cooper, R.L., Lamb-IV, J.C., Barlow, S.M., *et al*. (2006) A tiered approach to life stages testing for agricultural chemical safety assessment. Crit Rev Toxicol. 36: 69-98.

Creasy, D.M. (2003). Evaluation of Testicular Toxicology: A Synopsis and Discussion of the Recommendations Proposed by the Society of Toxicologic Pathology. Birth Defects Research (Part B) 68:408-415

Crofton, K. M. and L. P. Sheets (1989) Evaluation of sensory system function using reflex modification of the startle response. J. Am. Coll. Toxicol. 8: 199-211.

Crofton K. M., Dean K. F., Menache M. G., and R. Janssen. (1990) Trimethyltin effects on auditory function and cochlear morphology. Toxicol. Appl. Pharmacol. 105:123-132.

Crofton, K.M., Howard, J.L., Moser, V.C., Gill, M.W., Reiter, L.W., Tilson, H.A., MacPhail, R.C. (1991) Interlaboratory comparison of motor activity experiments: Implications for neurotoxicological assessments. Neurotoxicol. Teratol., 13:599-609

Crofton K. M., Peele D.B., and M.E. Stanton. (1993) Developmental neurotoxicity following neonatal exposure to 3,3'-iminodipropionitrile in the rat. Neurotoxicol. Teratol. 15:117-129.

Crofton K. M., Janssen R., Prazma J., Pulver S., and Jr. S. Barone. (1994a) The ototoxicity of 3,3'-iminopropionitrile: Functional and morphological evidence of cochlear damage. Hearing Research 80: 129-140.

Crofton K. M., Lassiter T. L., and C. S. Rebert. (1994b) Solvent-induced ototoxicity in rats: An atypical selective mid-frequency hearing deficit. Hearing Research; 80: 25-30.

Crofton, KM, Foss JA, Hass, U, Jensen, KF, Leven, ED, and Parker, SP (2007). Undertaking positive control studies as part of developmental neurotoxicity testing. A report from the ILSI Research Foundation/Risk Science Institute export working group on neurotoxicity endpoints.
www.elsevier.com/locate/neutera

CSTEE (Scientific Committee on Toxicity, Ecotoxicity and the Environment) (1999) Opinion on Human and Wildlife Health Effects of Endocrine Disrupting Chemicals, with Emphasis on Wildlife and on Ecotoxicity Test Methods. Expressed at the 8th CSTEE plenary meeting, Brussels, 4 March 1999. DG XXIV, Consumer Policy and Consumer Health Protection, European Commission.

D'Hooge, R., De Deyn, P.P. (2001) Applications of the Morris water maze in the study of learning and memory. Brain Res. Rev, 36:60-90.

Daston, G.P. (1994) Relationship Between Maternal and Developmental Toxicity, in C.A. Kimmel and J. Buelke-Sam, Developmental Toxicology, 2nd Ed., Raven Press, Ltd., New York.

Denenberg V.H. (1969) Open-field Behavior in the rat: What does it mean? Ann. N.Y. Acad. Sci. 159(3): 852-859.

Dent, M.P. (2007) Strengths and limitations of using repeat-dose toxicity studies to predict effects on fertility. Regul Toxicol Pharmacol. 48: 241-258.

DEPA (Danish Environmental Protection Agency) (1995) Male Reproductive Health and Environmental Chemicals with Estrogenic Effects. Ministry of Environment and Energy, Denmark, Miljøprojekt nr. 290, 1995 (ISSN 0105-3094/ISBN 87-7810-345-2) and IPCS-WHO 2001,

ECETOC (1983) Monograph No. 5. Identification and Assessment of the Effects of Chemicals on Reproduction and Development (reproductive Toxicology). Brussels December 1983) ECETOC. Guidance on Evaluation of Reproductive Toxicity Data. Monograph No. 31, Brussels, February 2002 (ISSN-0773-6347-31).

ECETOC (1992) (European Centre for Ecotoxicology and Toxicology of Chemicals (former E. Chemical Industry Ecology and Toxicology Centre) Technical Report No. 47. EC 7th Amendment: "Toxic to Reproduction" Guidance on Classification, Brussels, August 1992 (ISSN-0773-8072-47).

ECETOC (2002) Guidance on Evaluation of Reproductive Toxicity Data. Monograph No. 31, Brussels, February 2002 (ISSN-0773-6347-31).

ECETOC (2004) Influence of Maternal Toxicity in Studies on Developmental Toxicity Workshop Report No. 4, Brussels October 2004.

EHC (1984) Environmental Health Criteria 30. Principles for Evaluating Health Risks to Progeny Associated with Exposure to Chemicals During Pregnancy, WHO Geneva (ISBN 92 4 154090 7).

Elsner J. (1991) The tactile-kinesthetic system of rats as an animal model for minimal brain dysfunction. Arch. Toxicol. 65: 465-473.

Elsner, J., Suter, K., Ulbrich, B., and G. Schreiner. (1986) Testing Strategies in Behavioral Teratology: IV. Review and General Conclusions. Neurobehav. Toxicol. Teratol. 8: 585-590.

Elsner, J., Hodel, B., Suter, K., Oelke, D., Ulbrich, B., Schreiner, G., Cuomo, V., Gagiano, R., Rosengren, L.E., Karlsson, J.E., and K.G. Haglid. (1988) Detection Limits of Different Approaches in Behavioral Teratology, and Correlation of Effects With Neurochemical Parameters. Neurobehav. Toxicol. Teratol. 10: 155-167.

EU (2003) Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC) No. 1488/94 on Risk Assessment for Existing Substances. Available at: [http://ecb.jrc.it/]

File, S. E. 2001. Factors controlling measures of anxiety and responses to novelty in the mouse. Behav Brain Res 125:151-157.

Fix, A.S. and Garman, R.H. (2000) Practical aspects of neuropathology: A technical guide for working with the nervous system. *Toxicol. Pathol*. 28: 122-131.

Fleeman TL, Cappon GD, Chapin RE, and Hurtt ME. (2005) Effect of feed restriction during organogenesis on embryo-fetal development in the rat. Birth Defects Research (Part B) 74:442-449.

Francis, E.Z. (1990) Developmental neurotoxicity study and response to public and SAP comment, Washington D.C., US EPA, 1-20.

Francis, E. Z., Kimmel, C. A., and D. C. Rees. (1990) Workshop on the qualitative and quantitative comparability of human and animal developmental neurotoxicity: Summary and implications. Neurotoxicol. Teratol. 12: 285-292.

Gabriel, M., Kubota, Y., Sparenborg, S., Straube, K., and B. A. Vogt. (1991) Effects of cingulate cortical lesions on avoidance learning and training-induced unit activity in rabbits. Exp. Brain. Res. 86: 585-600.

Garman, R.H., A.S. Fix, B.S. Jortner, K.F. Jensen, J.F Hardisty, L. Claudio, and S. Ferenc. (2001) Methods to identify and characterize developmental neurotoxicity for human health risk assessment: neuropathology. *Environmental Health Perspectives* 109 (Suppl. 1):93-100.

Geller, A.M. and Zenick, H. (2005) Aging and the environment: a research framework. Environmental Health Perspectives 113(9): 1257-1262.

Gerber, G. J., and D. O'Shaughnessy. (1986) Comparison of the behavioral effects of neurotoxic and systemically toxic agents: How discriminatory are behavioral tests of neurotoxicity. Neurobehav. Toxicol. Teratol. 8: 1-12.

Globally Harmonized System of Classification and Labelling of Chemicals (GHS).

Goodlett, C. R., Kelly, S. J., and J. R. West. (1987) Early postnatal alcohol exposure that procedures high blood alcohol levels impairs development of spatial navigation learning. Psychobiology 15(1): 64-74.

Gray, L.E., Jr., Ostby, J., Monosson, E., and Kelce, W. R. (1999). Environmental antiandrogens: low doses of the fungicide vinclozolin alter sexual differentiation of the male rat. Toxicol Ind Health 15: 48-64.

Gray, L. E., Ostby, J., Furr, J., Wolf, C. J., Lambright, C., Park, L., Veeramachaneni, D. N., Wilson, V., Price, M., Hotchiss, A., Orlando, E., and L. Guillette. (2001) Effects of environmental antiandrogens on reproductive develeopment in experimental animals. Human Reproduction Update 7: 248-264.

Gårdlund, A. T., Archer, T., Danielsson, K., Danielsson, B., Frederiksson, A., Lindquist, N.-G., Lindström, H., and J. Luthman. (1991) Effects of prenatal exposure to tributyltin and trihexyltin on behavior in rats. Neurotoxicol. Teratol. 13: 99-105.

Hass, U. (1993) Neurobehavioural teratology of industrial chemicals. Effects of prenatal exposure to organic solvents on postnatal development and behaviour - validation and use of a screening test battery in laboratory rats, Ph.D-thesis. Copenhagen:National Institute of Occupational Health, 1-108.

Hass, U., Hansen, E. V., and G Østergaard. (1994a) Experimental studies in laboratory animals. In Hass U *et al*., (eds) Occcupational reproductive toxicity - Methods and testing strategy for hazard assessment of workplace chemicals. Nordic Council of Ministers and National Institute of Occupational Health, Denmark.

Hass, U., Lund, S. P., and J. Elsner. (1994b) Effects of prenatal exposure to N-methylpyrrolidone on postnatal development and behavior in rats. Neurotoxicol. Teratol. 16: 241-249.

Hass, U., Lund, S. P., Simonsen, L., and A. S. Fries. (1995). Effects of prenatal exposure to xylene on postnatal development and behavior in rats. Neurotoxicol. Teratol. 17: 341-349.

Heindel, J. J. (1999) Oocyte quantitation and ovarian histology. In: An evaluation and interpretation of reproductive endpoints for human health risk assessment. Eds: G. Daston and C. Kimmel, International Life Sciences Institute Press.

Hellwig, J., van Ravenzwaay, B., Mayer, M., and C. Gembardi. (2000) Pre-and postnatal oral toxicity of vinclozolin in Wistar and Long-Evans rats. Reg. Tox. 32: 42-50.

Holson, R.R. and Pearce, B. (1992) Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species. *Neurotox. Teratol*. 14: 221-228.

Hood, R.D. and Miller, D.B. (1997) Maternally Mediated Effects on Development, in Handbook of Developmental Toxicology.

Hughes, E. (1986) Evolution of a behavioural screen, Huntingdon Research Center.

Hughes, E. W., and A. K Palmer. (1986) An assessment of pre-weaning development and behaviour in safety evaluation studies. Department of Reproductive Toxicology HCR (Abstract).

ILSI (2003). Principles and Practices for Direct Dosing of Pre-weaning Mammals in Toxicity Testing and Research.. A report of the ILSI Risk Science Institute Expert Working Group on Direct Dosing of Pre-weaning Mammals in Toxicity Testing. ILSI Press, Washington DC.

IPCS (International Programme on Chemical Safety) (2001). Principles for Evaluating Health Risks to Reproduction Associated with Exposure to Chemicals, Environmental Health Criteria 225, WHO Geneva, 2001.

Ison, J.R. (1984) Reflex modification as an objective test for sensory processing following toxicant exposure. Neurobehav. Toxicol. Teratol., 6:437–445.

Jakobsen B. M., and O. Meyer (1989) Extrapolation from in vivo/in vitro Experiments to Human Beings in Embryo-Foetal Damage and Chemical Substances, Working Party Report. National Food Agency of Denmark, Ministry of Health, publication No. 181, May 1989 (ISBN 87-503-7881-3/ISSN 0903-9783).

Janer, G., Hakkert, B.C., Slob, W., Vermeire, T. and A.H. Piersma. (2007) A retrospective analysis of the two-generation study: What is the added value of the second generation? Reprod Toxicol. 24: 103-113.

Jensen K.F. and Catalano, S.M. (1998) Brain morphogenesis and developmental neurotoxicology. In: *Handbook of Developmental Neurotoxicology*, Slikker, Jr. W., Chang, L.W. (eds.) Academic Press, New York, pages 3-41.

Jones, B. J., and D. J. Roberts (1968) The quantitative measurement of motor incoordination in naive mice using an accelerating rotarod. Pharm. Pharmac. 20: 302-304.

Kawakami, T., Ishimura, R., Nohara, K., Takeda, K,. Tohyama, C. and S. Ohsako. (2006) Differential susceptibilities of Holtzman and Sprague- Dawley rats to fetal death and placental dysfunction induced by 2,3,7,8-teterachlorodibenzo-p-dioxin (TCDD) despite the identical primary structure of the aryl hydrocarbon receptor. Toxicol. Appl. Pharmacol. 212:224-236.

Kaplan, M., and S. D. Murphy. (1972) Effect of acrylamide on rotarod performance and sciatic nerve beta-glucuronidase of rats. Toxicol. Appl. Pharmacol. 22: 259-268.

Kelly, S. J., Goodlett, C. R., Hulsether, S. A., and J. R. West. (1988) Impaired spatial navigation in adult female but not adult male rats exposed to alcohol during the brain growth spurt. Behav. Brain Res. 27: 247-257.

Kimmel, C. A., and J. Buelke-Sam. (1985) Collaborative Behavioral Teratology Study: Background and overview. Neurobehav. Toxicol. Teratol. 7: 541-546.

Koch, M. (1999) The neurobiology of startle. Prog. Neurobiol. 59:107-128.

Korenbrot, C. C., Huhtaniemi, I. T., and R. I. Weiner. (1977) Preputial separation as an external sign of pubertal development in the male rat. Biol. Reprod. 17: 298-303.

Ladefoged, O., Hass, U., and L. Simonsen. (1989) Neurophysiological and behavioural effects of combined exposure to 2,5-hexanedione and acetone or ethanol in rats. Pharmacol. Toxicol. 65: 372-375.

Lamb, J.C. (1985) Reproductive toxicity testing: evaluating and developing new testing system. J. Am. Coll. Toxicol. 4(2).

Lanning *et al*. (2002). Toxicologic Pathology, 30(4) pp 507-520.

Latendresse, J.R., Warbrittion, A.R., Jonassen, H., And Creasy, D.M. (2002). Fixation of testes and eyes using a modified Davidson's fluid: comparison with Bouin's fluid and conventional Davidson's fluid. *Toxicol Pathol* 30, 524-533.

Leblond, C. P. and Y. Clermont (1952). Definition of the stages of the cycle of the seminiferous epithelium in the rat. Ann. N. Y. Acad. Sci. 55: 548-573.

Levine, S., and P. L. Broadhurst, (1963) Genetic and ontogenetic determinants of behaviour. II. Effects of infantile stimulation on adult emotionality and learning in selected strains of rats. J. Comp. Physiol. Psychol. : 56-65.

Levine, S., Haltmeyer, G. C., Karas, G. G., and V. H. Denenberg. (1967) Physiological and behavioural effects of infantile stimulation. Physiol. Behav. 2:32-37.

Lochry, E. A., and E. P. Riley. (1980) Retention of passive avoidance and T-maze escape in rats exposed to alcohol prenatally. Neurobehav. Toxicol. 2.

Lochry, E. A., Hoberman, A. M., and M. S. Christian. (1986) Standardization and application of behavioral teratology screens. Safety evaluation and regulation of chemicals 3, Homburger (ed): Basel, Karger, 49-61.

Lore, R., and H., Avis. (1970) Effects of auditory stimulation and litter size upon subsequent emotional behavior in the rat. Develop. Psychobiol. 2(4):212-215.

LST (1989) Embryo-Foetal Damage and Chemical Substances, Working Party Report. National Food Agency of Denmark, Ministry of Health, publication No. 181, May 1989 (ISBN 87-503-7881-3/ISSN 0903-9783).

MacPhail RC (1999) Recent developments in neurotoxicology. Inhal Toxicol. 11(6-7):519-21.

MARTA (Middle Atlantic Reproduction and Teratology Association) and MTA (Midwest Teratology Association) (1995) Historical Control Data (1992-1994) for Developmental and Reproductive Toxicity Studies Using the Crl:CD@BR Rat. Charles River Laboratories, Inc.

MARTA (Middle Atlantic Reproduction and Teratology Association) and MTA (Midwest Teratology Association) (1996) Historical Control Data (1992-1994) for Developmental and Reproductive l'oxicity Studies Using the New Zealand White Rabbit. Hazleton Research Products, Inc., Denver, P A.

McIntyre, B. S., Barlow, N. J., and P. M. D. Foster. (2001) Androgen-mediated development in male rat offspring exposed to flutamide in utero: performance and correlation of early postnatal changes in anogenital distance and nipple retention with malformations in androgen-dependent tissue. Toxicol. Sci. 62: 236-249.

McIntyre, B.S., Barlow, N.J., Wallace, D. G., Maness, S.C., Gaido, K.W., and Foster, P.M.D. (2000). Effects of in utero exposure to linuron on androgen dependent reproductive development in the male Crl:CD(SD)BR rat. Toxicol Appl Pharmacol 167: 87-99.

Meaney, M. J., Aitken, D. H., Bhatnager S., Van Berkel, C., and R. M. Sapolsky. (1988) Postnatal handling attenuates neuroendocrine, anatomical, and cognitive impairments related to the aged hippocampus. Science 238: 766-768.

Meaney, M. J., Aitken, D. H., Bhatnager, S., sand R. M. Sapolsky. (1991) Postnatal handling attenuates certain neuroendocrine, anatomical, and cognitive dysfunctions associated with aging in female rats. Neurobiol. Aging 12: 31-38.

Metzdorff SB, Dalgaard M, Christiansen S, Axelstad M, Hass U, Kiersgaard MK, Scholze M, Kortenkamp A, Vinggaard AM. (2007) Dysgenesis and histological changes of genitals and perturbations of gene expression in male rats after in utero exposure to antiandrogen mixtures. Toxicol Sci. 98(1):87-98.

Meyer, Otto A. Hugh A. Tislon, Weldon C. Byrd and Michael T. Riley (1979): "A Method for the Routine Assessment of Fore and Hindlimb Grip Strenght of Rats and Mice". Neurobehavioral Toxicology, 1, pp. 233-236.

Meyer, O., Jakobsen, B. M., and E. V. Hansen. (1989) Identification of embryo-foetal toxicity by means of animal studies. In: Andreasen, P. B., Brandt, N. J., Cohr, K.-H., Hansen, E. V., Hass, U., Hauge, M., Jakobsen, B.M., Knudsen, I., Lauritsen, J. G., Melchior, J. C., Meldgaard, L., Meyer, O., Olsen, J. H., Palludan, B., Poulsen, E. Embryofoetal damage and chemical substances. Working party report, Copenhagen: Levnedsmiddelstyrelsen, 1989, 1-127. (Translated version of Danish report from 1986).

Meyer, O and T. Svendsen (2003) Animal Models in Pharmacology and Toxicology. In: Handbook of Laboratory Animal Science, Second Edition, Vol II, Animal Models (Eds. J. Hau and G. L. Van Hoosier, Jr), CRC Press2003.

Miller, D. B., Eckerman, D. A., Krigman, M. R., and L. D. Grant. (1982) Chronic neonatal organotin exposure alters radial-arm maze performance in adult rats. Neurobehav. Toxicol. Teratol. 4:185-190.

Moser VC, Becking GC, MacPhail RC, Kulig BM (1997) The IPCS collaborative study on neurobehavioral screening methods. Fundam Appl Toxicol. 35(2):143-51.

Moser V.C., Walls I., Zoetis T. (2005) Direct dosing of preweaning rodents in toxicity testing and research: Deliberations of an ILSI RSI expert working group. International Journal of Toxicology 24(2): 87-94.

Moore, K.L. (1989) *Before We Are Born: Basic Embryology and Birth Defects.* 3rd edition. W.B. Saunders Co., Los Angeles.

Morris, R. (1981) Spatial Localization Does Not Require the Presence of Local Cues. Learning and Motivation 12: 329-260.

Morris, R. (1984) Developments of a water-maze procedure for studying spatial learning in the rat. J. Neuro. Sci. Meth. 11: 47-60.

Morris, R., Garrud, P., Rawlins, J. N. P., and J. O'Keefe. (1982) Place navigation impaired in rats with hippocampal lesions. Nature 297: 681-683.

Mylchreest, E., Sar, M., Catley, R. C., and P. M. D. Foster. (1999) Disruption of androgen-regulated reproductive development by di(n-butyl) phthalate during late gestation in rats is different from flutamide. Toxicol. Appl. Pharmacol. 156: 81-95.

Narotsky, M.G., Brownie, C.F., and Kavlock, R.J. (1997) Critical period of carbon tetrachloride-induced pregnancy loss in Fischer-344 rats, with insights into the detection of resorption sites by ammonium sulfide staining. Teratology 56: 252-61.

Nelson, S. and G. Gibori. (1993) Dispersion, separation, and culture of the different cell populations of the rat corpus luteum. In: Heindel, J.J. and R.E. Chapin (ed.), Methods in Toxicology, Volume 3, Part B, Female Reproductive Toxicology, Academic Press, Inc., London.

NIOH (AMI) (1994) Occupational reproductive toxicity. Methods and testing strategies for hazard assessment of workplace chemicals. Nordic Council of Ministers and National Institute of Occupational Health, Denmark. Copenhagen 1994 (ISBN: 87-7534-455-6).

Nordberg, M., Duffus, J.H., Templeton M. (2004). Glossary of terms used in toxicokinetics (IUPAC recommendations 2003). Pure Appl. Chem., Vol. 76, No. 5, 1033-1082.

Nunez, J. F., Ferré, P., García, E., Escorihuele, R. M., Fernández-Teruel, A., and A. Tobena. (1995) Postnatal handling reduces emotionality ratings and accelerates two-way active avoidance in female rats. Physiol. Behav. 831-835.

OECD (1983) Test Guideline 415. OECD Guideline for Testing of Chemicals. One-generation reproduction toxicity study. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (1984) Test Guideline 478. OECD Guideline for Testing of Chemicals. Genetic toxicology: rodent dominant lethal test. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (1995). Test Guideline 407. OECD Guideline for Testing of Chemicals. Repeated dose 28-day oral toxicity study in rodents. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (1995). Test Guideline 421. OECD Guideline for Testing of Chemicals. Reproduction/developmental toxicity screening test. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (1996). Test Guideline 422. OECD Guideline for Testing of Chemicals. Combine repeated dose toxicity study with the reproduction/developmental toxicity screening test. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (1998). Test Guideline 408. OECD Guideline for Testing of Chemicals. Repeated dose 90-day oral toxicity study in rodents. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (1998). Test Guideline 409. OECD Guideline for Testing of Chemicals. Repeated dose 90-day oral toxicity study in non-rodents. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

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. Available: [http://www.oecd.org/document/30/0,3343,en_2649_34377_1916638_1_1_1_1,00.html].

OECD (2000) Report of the OECD expert consultation meeting on developmental neurotoxicity testing, Washington, DC, October 23-25, 2000.

OECD (2001a). Test Guideline 414. OECD Guideline for Testing of Chemicals. Prenatal developmental toxicity study. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (2001b). Test Guideline 416. OECD Guideline for Testing of Chemicals. Two-generation reproduction toxicity study. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (2002a) IPCS (International Programme on Chemical Safety). Principles for Evaluating Health Risks to Reproduction Associated with Exposure to Chemicals, Environmental Health Criteria 225, WHO Geneva, 2001.

OECD (2002b). Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals. OECD Environment, Health and Safety Publications: Series on Testing and Assessment, No. 21. Available: [http://www.oecd.org/document/30/0,3343,en_2649_34377_1916638_1_1_1_1,00.html].

OECD (2004) Guidance document for neurotoxicity testing, OECD Series on Testing and Assessment, Number 20, Available: [http://www.oecd.org/document/30/0,3343,en_2649_34377_1916638_1_1_1_1,00.html].

OECD (2007a) Test Guideline 426. OECD Guideline for Testing of Chemicals. Developmental Neurotoxicity Study. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html].

OECD (2007b) Test Guideline 440. OECD Guideline for Testing of Chemicals. The Uterotrophic Bioassay: A short-term screening test for oestrogenic properties. Available: [http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html]

OECD(2008). Draft Guidance document for histopathologic evaluation of endocrine and reproductive tests. Available [http://www.oecd.org/document/12/0,3343,en_2649_34377_1898188_1_1_1_1,00.html]

OECD. Draft Test Guideline. OECD Guideline for Testing of Chemicals. The Hershberger assay.

OECD. Draft Test Guideline. OECD Guideline for Testing of Chemicals. Revised TG407.

Olton, D. S., Becker, J. T., and G. E. Handelmann. (1979) Hippocampus, space, and memory. Behav. Brain Sci. 2: 313-365.

Olton, D. S. (1987) The radial arm maze as a tool in behavioural pharmacology. Physiol Behav. 40:793-797.

Palmer, A.K. (1981) Regulatory requirements for reproductive toxicology: theory and practice. In: Developmental toxicology, (Eds.) Kimmel, C. A. and Buelke-Sam, J., New York, Raven Press, 259-288.

Palmer, A. K. (1986) A simpler multi-generation study. International Congress of Pesticide Chemistry; 1-20 (Abstract).

Peele, D.B., Allison, S.D., Crofton, K.M. (1990) Learning and memory deficits in rats following exposure to 3,3'-iminopropionitrile. Toxicol. Appl. Pharmacol., 105:321-332.

Perreault SD, Cancel AM (2001) Significance of incorporating measures of sperm production and function into rat toxicology studies. Reproduction 121(2):207-16.

Prophet, E.B., Mills, B., Arrington, J.B., Sobin, L.H. (1994) *Laboratory Methods in Histotechnology*, American Registry of Pathology, Washington, DC, pages 84-107.

Pryce CR and Feldon J (2003) Long-term neurobehavioural impact of the postnatal environment in rats: manipulations, effects and mediating mechanisms. Neurosci Biobehav Rev 27 (1-2), 57-71.

Pryor, G. T., Dickinson, J., Howd, R. A., and C. S. Rebert. (1983a) Neurobehavioral effects of subchronic exposure of weanling rats to toluene and hexane. Neurobehav. Toxicol. Teratol. 5: 47-52.

Pryor, G. T., Dickinson, J., Howd, R. A., and C. S. Rebert. (1983b) Transient cognitive deficits and high-frequency hearing loss in weanling rats exposed to toluene. Neurobehav. Toxicol. Teratol. 5: 53-57.

Raffaele, K.C., W. Sette, J.D. Doherty, S. Makris, and K.M. Crofton. (2005) Neuropathological findings in developmental neurotoxicity testing: comparison of qualitative and quantitative evaluations, poster presentation no. 977, 44th Annual meeting of the Society of Toxicology, New Orleans, LA, *The Toxicologist* 84(S-1):200.

Raimondo, S., and M. T. Draghetti. (1990) Influence of length of intra-uterine life on the appearance of developmental markers and neurobehavioural reflexes in rat. 18. Conference of the European Teratology Society 1990; (Abstract).

Reiter, L. W., and R. C. MacPhail. (1982) Factors influencing motor activity measurements in neurotoxicology. In Mitchell, C. T, (ed). Nervous system toxicology.

Rice, D. and Barone, S. Jr. (2000) Critical periods of vulnerability for the developing nervous system: Evidence from humans and animal models. *Environmental Health Perspectives* 108(Suppl. 3):511-533.

Robbins, T.W. (1977) A critique of the methods available for the measurement of spontaneous motor activity, Handbook of Psychopharmacology, Vol. 7, Iverson, L.L., Iverson, D.S., Snyder, S.H., (eds.) Plenum Press, New York, pp. 37-82.

Rosenthal, R., and K. L. Fode. (1963) The effect of experimenter bias on the performance of the albino rat. Behav. Science 8: 183-189.

Rosenzweig, M. R., and E. L. Bennett. (1969) Effects of differential environment on brain weights and enzyme activities in gerbils, rats, and mice. Develop. Psychobiol. 2: 87-95.

Ruppert PH, Dean KF, Reiter LW (1984) Neurobehavioral toxicity of triethyltin in rats as a function of age at postnatal exposure. Neurotoxicology 5(4):9-22.

Ruppert, P.H., Dean, K.F., Reiter, L.W. (1985) Development of locomotor activity of rat pups in figure-eight mazes. Dev. Psychobiol., 18:247-260

Russel, L. D., Ettlin, R. A., Sinha Hikim, A. P., and E.D. Clegg. (1990) Histological and Histopathological Evaluation of the Testis. Cache River Press.

Saillenfait, A. M., and B. Vannier. (1988) Methodological proposal in behavioural teratogenicity testing: assessment of propoxyphene, chlorpromazine, and vitamin A as positive controls. Teratology 37: 185-199.

Schapiro, S., Salas, M., and B. A. Shaywitz. (1970) Hormonal effects on the ontogeny of swimming ability in the rat: Assessment of central nervous system development. Science 193: 146-151.

Schardein, J. L., Petrere, J., A., Hentz, D. L., Camp, R. D. and Kurtz, S., M. (1978) Cannibalistic traits observed in rats treated with a teratogen. Laboratory Animals, 12, 81-83.

Schardein *et al*., (1985) Evaluation of Human Risk, section VII of Principles of Teratogenesis Applicable to Human Exposure to Drugs and Chemicals in Chemically induced Birth Defects, Marcel Dekker INC.

Shield, M.A. and Mirkes, P.E. (1998) Apoptosis. In Slikker, W and Chang, LW (eds.), *Handbook of Developmental Neurotoxicology*, Academic Press, New York, pages 159-188.

Schiorring, E. (1979) An open field study of stereotyped locomotor activity in amphetamine treated rats. Psychopharmacol. 66: 281.

Schröder, H.J. (1995) Comparative aspects of placental exchange functions. Eur. J. Obstet. Gynecol. Reprod. Biol. 63:81-90. Takayama, S., Akaike, M., Kawashima, K., *et al*. (1995) A collaborative study in Japan on optimal treatment period and parameters for detection of male fertility disorders induced by drugs in rats. J Am Col Toxicol. 14: 266-292.

Seed, J., Chapin, R. E., Clegg, E. D., Darney, S. P., Dostal, L., Foote, R. H., Hurtt, M. E., Klinefelter, G. R., Makris, S., Schrader, S., Seyler, D., Sprando, R., Treinen, K. A., Veeramachaneni, R. and L. D. Wise. (1996). Methods for assessing sperm motility, morphology, and counts in the rat, rabbit, and dog: A consensus report. Reprod. Toxicol. 10:237-244.

Selevan, S. G., Kimmel, C. A., and P. Mendola. (2000) Identifying critical windows of exposure for children's health. Environmental Health Perspectives 108(3):451-455.

Solecki, R., Bürgin, H., Buschmann, J., Clark, R., Duverger, M., Fialkowski, O., Guittin, P., Hazelden, K.P., Hellwig, J., Hoffmann, E., Hofmann, T., Hübel, U., Khalil, S., Lingk, W., Mantovani, A., Moxon, M., Müller, S., Parkinson, M., Paul, M., Paumgartten, F., Pfeil, R., Platzek, T., Rauch-Ernst, M., Scheevelenbos, A., Seed, J., Talsness, C.E., Yasuda, M., Younes, M., and Chahoud, I. (2001) Harmonization of rat fetal skeletal terminology and classification. Report of the Third Workshop on the Terminology in Developmental Toxicology, Berlin, 14-16 September 2000. Reprod Toxicol. 15(6): 713-21.

Solecki, R., Bergmann, B., Bürgin, H., Buschmann, J., Clark, R., Druga, A., Van Duijnhoven, A.J., Duverger, M., Edwards, J., Freudenberger, H., Guittin, P., Hakaite, P., Heinrich-Hirsch, B., Hellwig, J., Hofmann, T., Hübel, U., Khalil, S., Klaus, A, Kudicke, S., Lingk, W., Meredith, T., Moxon, M., Müller, S., Paul, M., Paumgartten, F., Röhrdanz, E., Pfeil, R., Rauch-Ernst, M., Seed, J., Spezia, F., Vickers, C., Woelffel, B., and Chahoud, I. (2003) Harmonization of rat fetal external and visceral terminology and classification. Report of the Fourth Workshop on the Terminology in Developmental Toxicology, Berlin, 18-20 April 2002. Reprod Toxicol. 17(5): 625-637.

Spark, C. and Dawson, A. B. (1928). The order and time of appearance of centers of ossification in the fore and hind limbs of the albino rat, with special reference to the possible influence of the sex factor. American Journal of Anatomy 41, 411-445.

Stanton, M. E. and J. H. Freeman. (1994) Eyeblink conditioning in the infant rat: An animal model of learning in developmental neurotoxicology. Environ. Health Perspect. 102: 131-139.

Staples, R. E. (1974) Detection of visceral alterations in mammalian fetuses. Teratology 9(3):A37- A38.

Tamura, R. N., and J. Buelke-Sam. (1992) The use of repeated measures analyses in developmental toxicology studies. Neurotoxicol. Teratol. 14: 205-210.

Tilson, H.A. (1994) Developmental neurotoxicology risk assessment. In: Harry, G.J. (ed.) *Developmental Neurotoxicology.* CRC Press, Boca Raton, pages 157-169.

Toppari, J., Larsen, J., Christiansen, P., Giwercman, A., Grandjean, P., Guillette Jr. L. J., Jegou, B., Jensen, T. K., Jouannet, P., Keiding, N., Leffers, H., McLachlan, J. A., Meyer, O., Muller, J., Rajpert-De Meyts, E., Scheike, T., Sharpe, R., Sumpter, J., and N. E. Skakkebaek. (1996) Male Reproductive Health and Environmental Xenoestrogens. Environmental Health Perspectives 104(Suppl 4):741-803.

United Nations (UN) (2003) Globally Harmonized System of Classification and Labelling of Chemicals (GHS). ST/SG/AC.10/30, UN New York and Geneva. Available:
[http://www.unece.org/trans/danger/publi/ghs/officialtext.html]

US EPA. (1990) Developmental neurotoxicity study and response to public and SAP comment. Washington D.C.

US EPA (1991) Guidelines for developmental toxicity risk assessment. Federal Register 56(234):63798-63826.

US EPA. (1996) Guidelines for Reproductive Toxicity Risk Assessment; October 1996. EPA/630/R-96/009.

US EPA (1998a) Health Effects Test Guidelines OPPTS 870.6300 Developmental Neurotoxicity Study. Available:
[http://www.epa.gov/opptsfrs/OPPTS_Harmonized/870_Health_Effects_Test_Guidelines/Series/870-6300.pdf]

US EPA (1998b) Guidelines for Neurotoxicity Risk Assessment. US EPA 630/R-95/001F. Available: [http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?PrintVersion=True&deid=12479].

US EPA (2002) A review of the reference dose and reference concentration processes. Risk Assessment Forum, EPA/630/P-02/002F, Washington, D.C., available at: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=55365.

US National Toxicology Program. (2006) C. elegans: A Medium Throughput Screening Tool for Toxicology. Website http://ntp.niehs.nih.gov/files/WormToxFS06.pdf (Accessed 12.1.2008).

Vorhees, C. V. (1983) Foetal anticonvulsant syndrome in rats: Dose and period response relationships of prenatal diphenylhydantoin, trimethadione, and phenobarbital exposure on the structural and functional development of the offspring. Pharmacol. Exp. Ther. 227: 274-287.

Vorhees, C. V. (1985) Comparison of the collaborative behavioral teratology study and Cincinnati behavioral test batteries. Neurobehav. Toxicol. Teratol. 7: 625-633.

Vorhees, C.V. (1987) Maze learning in rats: A comparison of performance in two water mazes in progeny prenatally exposed to different doses of phenytoin. Neurotoxicol. Teratol., 9:235-241.

Vorhees, C. V., Weisenburger, W. P., Acuff-Smith, K. D., and R. D. Minck. (1991) An analysis of factors influencing complex watermaze learning in rats: effects of task complexity, path order and escape assistance on performance following prenatal exposure to phenytoin. Neurotoxicol. Teratol.; 13: 213-222.

Walsh RN, Cummins RA (1976) The open-field test: a critical review. Psychol Bull 83: 482-504

Walsh, T. J., ad J. J. Chrobak. (1987) The use of radial arm maze in neurotoxicology. Physiol. Behav. 40: 799-803.

Weisenburger, W. P., Minck, D. R., Acuff, K. D., and C. V. Vorhees. (1990) Working memory and the Olton radial-arm maze: Effects of maternal phenylketonuria and prenatal phenytoin in rats. Teratol. 41: 615.

Wier, P. J., Guerriero, F. J., and R. F. Walker. (1989) Implementation of a primary screen for developmental neurotoxicity. Fundam. Appl. Toxicol. 13:118-136.

Wilson, J. D. (1965) Embryological Considerations in Teratology. In: Teratology: Principles and Techniques, Chapter 10. (Eds.) J. G. Wilson and J. Warkany, Chicago: The University of Chicago Press, 256.

Wise, L. D., Beck, S. L., Beltrame, D., Beyer, B. K., Chahoud, I., Clark, R. L., Clark, R., Druga, A. M., Fueston, M. H., Guittin, P., Henwood, S. M., Kimmel, C. A., Lindstrom, P., Palmer, A. K., Petrere, J. A., Solomon, H. M., Yasuda, M., and R. G. York. (1997) Terminology of developmental abnormalities in common laboratory mammals (version 1). Teratol.55: 249-292.

APPENDIX I

Definitions

Abortion: the premature expulsion from the uterus of the products of conception: of the embryo or of a nonviable foetus (OECD, 2001a).

Anomalies: structural alterations in development that include both malformations and variations (Wise *et al*., 1997).

Conceptus: the sum of derivatives of a fertilised ovum at any stage of development from fertilisation until birth including the extra-embryonic membranes as well as the embryo or foetus (OECD, 2001a).

Developmental toxicology: the study of adverse effects on the developing organism that may result from exposure prior to conception, during prenatal development, or postnatally to the time of sexual maturation. The major manifestations of developmental toxicity include 1) death of the organism, 2) structural abnormality, 3) altered growth, and 4) functional deficiency(OECD, 2001a).

Embryo: the early or developing stage of any organism, especially the developing product of fertilisation of an egg after the long axis appears and until all major structures are present (OECD, 2001a).

Embryotoxicity: detrimental to the normal structure, development, growth, and/or viability of an embryo (OECD, 2001a).

Fertility: The capacity to become pregnant or to impregnate. In humans, the capacity for producing viable offspring (IPCS, 2001).

Fertilization: The fusion of the sperm and ovum resulting in the restoration of the diploid number of chromosomes (IPCS, 2001).

Foetus: the unborn offspring in the post-embryonic period (OECD, 2001a).

Foetotoxicity: detrimental to the normal structure, development, growth, and/or viability of a foetus (OECD, 2001a).

Gestation: Length of time between conception and birth (IPCS, 2001).

Hazard characterisation: Involves determining whether or not an agent poses a hazard, at what doses and under what conditions of exposure (IPCS, 2001).

Hazard identification: The identification of the inherent capability of a substance to cause adverse effects (IPCS, 2001).

Implantation (nidation): attachment of the blastocyst to the epithelial lining of the uterus, including its penetration through the uterine epithelium, and its embedding in the endometrium (OECD, 2001a).

Malformation/Major Abnormality: Structural change considered detrimental to the animal (may also

be lethal) and is usually rare (Wise at al., 1997).

No-Observed-Adverse-Effect-Level (NOAEL): Highest concentration or amount of a substance, found by experiment or observation, that causes no detectable adverse alteration of morphology, functional capacity, growth, development or life span of the target organism under defined conditions of exposure (Nordberg *et al*., 2004).

Perinatal: The period around the time of birth.

Pregnancy: The condition of having an implanted embryo or fetus in the body, after fusion of an ovum and spermatozoon (IPCS, 2001).

Reproductive toxicity: includes adverse effects on sexual function and fertility in adult males and females as well as developmental toxicity in the offspring (GHS).

Resorption: a conceptus which, having implanted in the uterus, subsequently died and is being, or has been resorbed (OECD, 2001a):

Early resorption: evidence of implantation without recognisable embryo/foetus.

Late resorption: dead embryo or foetus with external degenerative changes.

Risk assessment: An empirical based paradigm that estimates the risk of adverse effect(s) from exposure of an individual or population to a chemical, physical or biological agent. It includes the components of hazard identification, assessment of dose-response relationships, exposure assessment and risk characterization (IPCS, 2001).

Sexual maturation: Achievement of full development of sexual function and reproductive system (IPCS, 2001).

Teratogenicity: Induction of structural abnormality (IPCS, 2001).

Variation/Minor Abnormality Structural change considered to have little or no detrimental effect on the animal; may be transient and may occur relatively frequently in the control population (Wise, 1997).

APPENDIX 2

Regulatory use of data from reproductive toxicity studies – Classification and Labelling

Data from reproductive toxicity studies may be used in Hazard Classification and Labelling. The following are examples of some of the classification and labeling systems in use.

United Nations

The United Nations (UN) Globally Harmonized System of Classification and Labelling of Chemicals (GHS) (UN 2003) provide a harmonized classification system for reproductive toxicity where chemical substances are allocated to one of two categories. The GHS is based on regulatory requirements in the US and Canada (workplace, consumers and pesticides), the European Union directives for classification and labelling of substances and preparations, and the UN Recommendations on the Transport of Dangerous Goods. Effects on reproductive ability or capacity, and on development, are considered as separate issues. In addition, effects on lactation are allocated to a separate hazard category. The two categories are:

- Category 1: Known or presumed human reproductive toxicant.
- Category 1A: Known to have produced an adverse effect on reproductive ability or capacity or on development in human. (The placing of the substance is largely based on evidence from humans)
- Category 1B: Presumed to produce an adverse effect on reproductive ability or capacity or on development in humans. (The placing of a substance in this category is largely based on evidence from experimental animals)
- Category 2: Suspected human reproductive or developmental toxicant (This category includes substances for which there is some evidence from humans or experimental animals, possibly supplemented with other information of an adverse effect on reproductive ability or capacity, or on development, in the absence of other toxic effects, or if occurring together with other toxic effects the adverse effect on reproduction is considered not to be a secondary non-specific consequence of the other toxic effects, and where the evidence is not sufficiently convincing to place the substance in category 1)

European Union

The hazard-based classification system used by all EU Member Countries is defined in Annex VI to Commission Directive 67/548/EEC (Anon 1967), last updated by Commission Directive 2001/59/EC (Anon 2001). In the EU system reproductive toxicity is divided into effects on fertility and developmental toxicity. In addition, substances which may interfere with lactation or are present in the breast milk in significant amounts can be classified as hazardous to breast fed babies. There are three categories:

- Category 1: This category includes: (i), substances known to impair fertility in humans; and (ii), substances known to cause developmental toxicity in humans. They are placed in this category if there is sufficient evidence to establish a causal relationship between human exposure to the substance and impaired fertility and if there is sufficient evidence to establish a causal relationship between human exposure to the substance and subsequent developmental toxic effects in the progeny.

- Category 2: This category includes: (i) substances which should be regarded as if they impair fertility in humans, and (ii), substances which should be regarded as if they cause developmental toxicity in humans. They are assigned to this category if there is sufficient evidence to provide a strong presumption that human exposure to the substance may result in (a), impaired fertility on the

basis of: clear evidence in animal studies of impaired fertility in the absence of toxic effects, or, evidence of impaired fertility occurring at around the same dose levels as other toxic effects but which is not a secondary non-specific consequence of the other toxic effects; other relevant information and (b), developmental toxicity, generally on the basis of: clear results in appropriate animal studies where effects have been observed in the absence of signs of marked maternal toxicity, or at around the same dose levels as other toxic effects but which are not a secondary non-specific consequence of the other toxic effects; other relevant information.

- Category 3: This category includes: (i), substances which cause concern for human fertility; and (ii), substances which cause concern for humans owing to possible developmental toxic effects. They are assigned to this category generally on the basis of results in appropriate animal studies which provide sufficient evidence to cause a strong suspicion of (a), impaired fertility in the absence of toxic effects, or evidence of impaired fertility occurring at around the same dose levels as other toxic effects, but which is not a secondary non-specific consequence of the other toxic effects, but where the evidence is insufficient to place the substance in category 2, and/or other relevant information and (b), developmental toxicity in the absence of marked maternal toxicity, or at around the same dose levels as other toxic effects but which are not a secondary non-specific consequence of the other toxic effects, but where the evidence is insufficient to place the substance in category and/or other relevant information.

Essentially all European countries which operate a classification system for reproductive toxicity are currently using or are likely to adopt the EU system in the near future. Guidance for hazard classification can be found in the EU Technical Guidance Document on Risk Assessment (EU 2003).

APPENDIX 3

Overall approach to testing

The objective is to give guidance on an approach to hazard characterization with regard to reproductive toxicity. The guidance seeks to ensure that these data are obtained in the most efficient and humane manner so that animal usage and costs are minimised. This is an iterative process involving the consideration of all available information and is conducted on a case-by-case basis.

A stepwise assessment/testing strategy is recommended. To minimise animal usage and optimise allocation of resources, data should be assessed following each step of testing to decide if they are adequate for the evaluation of the risk arising from the intended use of the chemical and from unintended or environmental exposure, or if further testing is needed.

Useful information can be derived from physical/chemical properties, in vitro studies or structure-activity-relationship (Q)SAR, mode of action information or in vivo animal studies and epidemiological data. A variety of toxicological data for chemicals may be available including studies on acute, sub-chronic, chronic toxicity, genotoxicity, toxicokinetics, toxicodynamics, immunotoxicity, neurotoxicity, reproductive toxicity and prenatal developmental toxicity.

QSAR data: The current knowledge concerning QSAR's in reproductive toxicity is limited, and the absence of structural alerts should not preclude testing a substance. However, when there is some structural relationship between the chemical and other chemicals shown to cause toxic effect on reproduction, such information should be considered and may trigger testing.

Epidemiological data**:** It is very difficult unequivocally to associate human exposure to a specific substance with adverse effects on reproduction. It is possible when the adverse effect is a rare birth defect and the exposure is very well characterised. However, it is difficult to identify a cause-effect relationship for reproductive effects which have a high "natural" incidence (e.g., spontaneous abortion) and/or which may be predispose to recall bias. Also, some effects can be very subjective (e.g., effects on libido) and there may be mixed exposures and/or lifestyle related confounding factors. Nevertheless, well designed and well reported epidemiological studies in which both reproductive and relevant non-reproductive effects are described will contribute to the weight of evidence for whether or not the substance is toxic to reproduction

If there is clear evidence of reproductive toxic effects in humans, the chemical should be considered toxic to reproduction and further studies may not be needed for hazard characterisation.

If there is some evidence of reproductive toxic effects in humans, experimental studies including toxicokinetics to investigate the causal relationship between exposure to a chemical and the effect may be useful.

When there is some evidence of developmental toxicity in humans, a prenatal developmental toxicity study or developmental neurotoxicity study in experimental animals may be useful, depending on the effect seen in humans. If the effects observed in humans indicate endocrine disruption, effects on development of sex organs or postnatally induced developmental toxicity, a two-generation reproductive toxicity study could provide further relevant information.

When there is some evidence of impaired fertility or indications of reproductive toxicity in humans, such as increases in sperm anomalies or menstrual disturbances, a one- or multi-generation study may be useful.

Toxicological data in experimental animals

In the following, suggestions will be given for possible further testing for hazard characterisation based upon the data from OECD reproductive toxicity tests or toxicity data.

Relative Sensitivity

The relative sensitivity of pregnant and non-pregnant animals can be judged by comparing data from repeated dose toxicity studies to data from reproductive toxicity studies. However, this assessment has inherent problems, since the parameters on systemic toxicity measured in the maternal animals in reproductive toxicity tests are generally limited compared to those assessed in non-pregnant females in repeated dose toxicity studies.

In order to have a sufficient background to determine the sensitivity for exposure during the developmental period compared to adulthood there is a need for studies where end-points are investigated similarly for both age groups. This could be a two-generation study incorporating developmental neurotoxicity end points and similar investigations of systemic effects in offspring and adults as is included in repeat dose toxicity studies.

Results from One or Two-Generation Studies

1. Possible effects on fertility

a) *Clear evidence of effects:* When there is clear evidence of effects on fertility in animals, further studies may not be needed for hazard characterisation. However, further studies may provide information for dose-effects relationship and the relevance for humans.

b) *Equivocal evidence:* Equivocal effects on fertility could suggest the necessity of mechanistic studies or a further study with a modified design. The study should be designed so that it will be possible to determine whether an effect is mediated via the male or the female or caused by general toxicity. In some cases specific studies e.g., a continuous breeding study could be relevant.

c) *Clear evidence of no effects*: If there is no evidence of effects on male or female fertility, sperm quality or oestrus cyclicity further studies are normally not needed for the hazard assessment of the chemical. However, if there is some, but limited evidence of fertility effects in humans further studies may be needed e.g., in another laboratory animal species or including of additional endpoints.

2. Possible effects on *in utero* development

a) *Clear evidence of effects:* Clear evidence of developmental effects e.g., decreased litter size, stillbirth, anomalies, postnatal death or functional anomalies will generally be sufficient for hazard identification purposes. However, further data on toxicodynamics, metabolism and/or toxicokinetics may necessary to understand the relevance for humans.

b) *Equivocal evidence:* Equivocal evidence of developmental toxicity may be further investigated in a prenatal developmental toxicity study or a developmental neurotoxicity study.

c) *Clear evidence of no effects: G*eneration-type studies have limitations in detecting developmental toxicity. Thus, the lack of evidence of any developmental toxicity is not sufficient to exclude the possibility of developmental toxicity.

Results from Prenatal Developmental Toxicity Studies

Possible effects on *in utero* development

a) *Clear evidence of effects:* Clear evidence of developmental effects will generally be sufficient for hazard identification purposes. However, further data on toxicodynamics, metabolism and/or toxicokinetics may necessary to understand the relevance for humans.

b) Equivocal evidence: Equivocal results could suggest the necessity of mechanistic studies or a further study with modified design or using another species. If developmental toxicity has only been observed at doses causing marked toxic effects on the mother, studies using lower/less spaced dose levels and looking for specific endpoints are recommended. In some cases, specific studies of developmental neurotoxicity could be relevant, e.g., if there are anomalies in the brain or if other information indicates a neurotoxic potential of the chemical in question.

c) Clear evidence of no effects: If there is no evidence of developmental toxicity, further studies are normally not needed for hazard assessment. However, if there is some, but limited evidence of developmental toxicity effects in humans, further studies may be needed including studies in another laboratory animal species. In some cases, specific studies of developmental neurotoxicity could be relevant, e.g., if other information indicates a neurotoxic potential of the chemical in question.

Results from Screening Tests (TG 421 and 422)

a) *Clear evidence of effects:* Clear evidence of effects may be sufficient for hazard identification purposes. However, further data on toxicodynamics, metabolism and/or toxicokinetics may necessary to understand the relevance for humans.

b) Equivocal evidence: Equivocal results could suggest the necessity of a more comprehensive study (e.g. TG 414, TG415, TG 416, TG 426), mechanistic studies or a further study with modified design or using another species.

c) Clear evidence of no effect. If there is no evidence of any effects further studies are normally not needed for hazard assessment. However, negative studies may insufficient to exclude a possible effect. The study design does not cover all relevant parameters (e.g., mating behaviour, functional effects, significant post-natal development) and has other limitation (e.g., relatively small numbers of animals per dose level or relatively short study duration), which means it may not provide sufficient data for a comprehensive assessment for reproductive toxicity. Therefore, it is recommended that where exposure considerations or structural alerts suggest that the risk assessment will ultimately require a robust investigation of reproductive endpoints, then TG 421 and 422 should be omitted and a more comprehensive study design (e.g. TG 414, TG 415, TG 416, TG 426) considered.

Results from Developmental Neurotoxicity Study (TG 426)

a) *Clear evidence of effects:* Clear evidence of developmental neurotoxicity in animals no further studies will usually be needed for hazard assessment. However, further data on toxicodynamics, metabolism and/or toxicokinetics may necessary to understand the relevance for humans.

b) Equivocal evidence: Substances giving rise to less clear manifestations of developmental neurotoxicity (such as retardation of postnatal reflex ontogeny) or no firm conclusion can be drawn from the study about the developmental neurotoxicity of the substance should be considered for further investigations so that the potential for induction of serious functional effects can be evaluated. Effects could be investigated further by additional testing, normally in the same species, with a different study

design or with larger group sizes and/or with inclusion of additional end-points.

c) Clear evidence of no effect: If there is no evidence of developmental neurotoxicity of the chemical, further studies are normally not needed for the hazard assessment of developmental neurotoxicity. However, further studies using another species, direct dosing of the offspring and/or examination of developmental neurotoxicity end points not covered yet should be considered when there is potential widespread exposure of women of childbearing age and indications of developmental neurotoxicity in humans.

Other Data from Experimental Animals

Results of Repeated Dose Toxicity Studies

Data from repeated dose toxicity studies in which there are marked adverse effects on the reproductive organs may provide sufficient evidence of effects to reproduction for the purpose of classification and labeling. In this case no further studies may be necessary. However, further data on toxicodynamics, metabolism and/or toxicokinetics may necessary to understand the relevance for humans. Data from repeated dose toxicity studies are not sufficient to exclude a possible effect on reproduction, since fertility may be impaired in the absence of any histological damage to the gonads.

If repeated dose toxicity studies have indicated testicular effects, a one- or two-generation study focusing on the development of the male sexual organs and fertility could provide useful further information. Additional group(s) in which only the males are dosed could be included. If QSAR indicates an effect on fertility, a specific mating study including sperm analysis might be considered as sufficient for the hazard evaluation.

If repeated dose toxicity studies have indicated ovarian effects, a one- or two-generation study focusing on parameters of female reproductive toxicity including oestrus cyclicity may provide useful further information. Additional group(s) in which only the females are dosed could provide additional information. Information from QSAR may further identify relevant parameters for inclusion in the study design.

If there is evidence of neurotoxic effect in adult animals, a developmental neurotoxicity study should be considered. Developmental neurotoxicity testing may be conducted to further characterise neurological effects in offspring observed in other studies, and should be considered if the substance has been shown to cause e.g. neuropathology or neurotoxicity in adults, to be a hormonally-active material *in vivo* (e.g., pituitary, thyroid, sex hormones), or to cause other types of toxicity, suggestive of nervous system involvement at a developmental stage

Other Data (In Vitro Tests, Kinetics/Dynamics)

If *in vitro* studies have indicated a potential for developmental toxicity, and QSAR and kinetic data support the relevance of this, further studies may not be needed for an initial hazard assessment. However, a prenatal developmental toxicity study would provide useful information. Generally, the predictive value of *in vitro* effects in isolated tissues and their relevance for human is not well understood. However, in vitro testing provides valuable information for the detection of endocrine disrupters. In particular, Brian *et al.* (2005) and Metzdorff *et al.* (2007) showed in vitro estrogenic multicomponent mixtures to be correctly predicted in vivo in both fish and rat model.

Chemicals shown to be a vitamin or hormone analogue or causing e.g., accumulation in the foetus, competitive metabolism, lowered blood pressure/disturbed circulation, lowered oxygen tension, formation of carboxyhemoglobin or genotoxic effects may be potential reproductive toxicants. For such chemicals, developmental toxicity studies should be considered as a first tier.

In addition, for chemicals shown to be a hormone analogue, causing hormonal effects in specific tests (e.g., Uterotrophic test, Hershberger test, receptor binding) or accumulation in reproductive organs, a two-generation reproductive toxicity study is usually appropriate.

Kinetic data (e.g. transfer of material to the fetus in utero or to the pup in lactation) are useful in the understanding and evaluation of data from the studies and may support appropriate dose level selection or efforts to modify exposures of animal on test.